

HANDBOOKS

IN ECONOMICS

Gilles Duranton
J. Vernon Henderson
William Strange

Regional and Urban Economics

VOLUME 5B

NORTH-HOLLAND

Handbook of
**REGIONAL AND
URBAN ECONOMICS**

This page intentionally left blank

Handbook of **REGIONAL AND URBAN ECONOMICS**

Volume 5B

Edited by

GILLES DURANTON

Wharton School, University of Pennsylvania,
Philadelphia, PA, USA, and CEPR

J. VERNON HENDERSON

Department of Geography, London School of Economics,
London, UK

WILLIAM C. STRANGE

Rotman School of Management, University of Toronto,
Toronto, ON, Canada



Amsterdam • Boston • Heidelberg • London • New York • Oxford
Paris • San Diego • San Francisco • Singapore • Sydney • Tokyo
North-Holland is an imprint of Elsevier



North-Holland is an imprint of Elsevier

Radarweg 29, PO Box 211, 1000 AE Amsterdam, The Netherlands
The Boulevard, Langford Lane, Kidlington, Oxford OX5 1GB, UK

Copyright © 2015 Elsevier B.V. All rights reserved.

Chapter 15, How Mortgage Finance Affects the Urban Landscape, Copyright © 2015 Elsevier B.V. and FRBNY. All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system or transmitted in any form or by any means electronic, mechanical, photocopying, recording or otherwise without the prior written permission of the publisher

Permissions may be sought directly from Elsevier's Science & Technology Rights Department in Oxford, UK: phone (+44) (0) 1865843830; fax (+44) (0) 1865 853333; email: permissions@elsevier.com.

Alternatively you can submit your request online by visiting the Elsevier web site at <http://elsevier.com/locate/permissions>, and selecting Obtaining permission to use Elsevier material.

Notices

Knowledge and best practice in this field are constantly changing. As new research and experience broaden our understanding, changes in research methods, professional practices, or medical treatment may become necessary.

Practitioners and researchers must always rely on their own experience and knowledge in evaluating and using any information, methods, compounds, or experiments described herein. In using such information or methods they should be mindful of their own safety and the safety of others, including parties for whom they have a professional responsibility.

To the fullest extent of the law, neither the Publisher nor the authors, contributors, or editors, assume any liability for any injury and/or damage to persons or property as a matter of products liability, negligence or otherwise, or from any use or operation of any methods, products, instructions, or ideas contained in the material herein.

British Library Cataloguing-in-Publication Data

A catalogue record for this book is available from the British Library

Library of Congress Cataloging-in-Publication Data

A catalog record for this book is available from the Library of Congress

ISBN: 978-0-444-59517-1 (Vol. 5A)

ISBN: 978-0-444-59531-7 (Vol. 5B)

For information on all North-Holland publications
visit our website at <http://store.elsevier.com/>



Working together
to grow libraries in
developing countries

www.elsevier.com • www.bookaid.org

Typeset by SPi Global, India

Printed and bound in the UK

Publisher: Nikki Levy

Acquisition Editor: J. Scott Bentley

Editorial Project Manager: Joslyn Chaiprasert-Paguio

Production Project Manager: Nicky Carter

Designer: Alan Studholme

INTRODUCTION TO THE SERIES

The aim of the *Handbooks in Economics* series is to produce Handbooks for various branches of economics, each of which is a definitive source, reference, and teaching supplement for use by professional researchers and advanced graduate students. Each Handbook provides self-contained surveys of the current state of a branch of economics in the form of chapters prepared by leading specialists on various aspects of this branch of economics. These surveys summarize not only received results but also newer developments from recent journal articles and discussion papers. Some original material is also included, but the main goal is to provide comprehensive and accessible surveys. The Handbooks are intended to provide not only useful reference volumes for professional collections but also possible supplementary readings for advanced courses for graduate students in economics.

Kenneth J. Arrow and Michael D. Intriligator

This page intentionally left blank

CONTENTS

<i>Foreword</i>	xv
<i>Contributors</i>	xvii

Volume 5A

Section I. Empirical Methods	1
1. Causal Inference in Urban and Regional Economics	3
Nathaniel Baum-Snow, Fernando Ferreira	
1.1. Introduction	4
1.2. A Framework for Empirical Investigation	6
1.3. Spatial Aggregation	20
1.4. Selection on Observables	23
1.5. IV Estimators	43
1.6. Regression Discontinuity	53
1.7. Conclusion	62
References	63
2. Structural Estimation in Urban Economics	69
Thomas J. Holmes, Holger Sieg	
2.1. An Introduction to Structural Estimation	70
2.2. Revealed Preference Models of Residential Choice	74
2.3. Fiscal Competition and Public Good Provision	79
2.4. The Allocation of Economic Activity Across Space	96
2.5. Conclusions	110
Acknowledgments	111
References	111
3. Spatial Methods	115
Steve Gibbons, Henry G. Overman, Eleonora Patacchini	
3.1. Introduction	116
3.2. Nonrandomness in Spatial Data	120
3.3. Spatial Models	124
3.4. Identification	136
3.5. Treatment Effects When Individual Outcomes Are (Spatially) Dependent	152
3.6. Conclusions	157

Appendix A: Biases with Omitted Spatial Variables	158
Appendix B: Hypothetical RCT Experiments for Identifying Parameters in the Presence of Interactions Within Spatial Clusters	161
References	164
Section II. Agglomeration and Urban Spatial Structure	169
4. Agglomeration Theory with Heterogeneous Agents	171
Kristian Behrens, Frédéric Robert-Nicoud	
4.1. Introduction	172
4.2. Four Causes and Two Moments: A Glimpse at the Data	175
4.3. Agglomeration	187
4.4. Sorting and Selection	211
4.5. Inequality	234
4.6. Conclusions	239
Acknowledgments	240
References	241
5. The Empirics of Agglomeration Economies	247
Pierre-Philippe Combes, Laurent Gobillon	
5.1. Introduction	248
5.2. Mechanisms and Corresponding Specifications	252
5.3. Local Determinants of Agglomeration Effects	270
5.4. Estimation Strategy	282
5.5. Magnitudes for the Effects of Local Determinants of Productivity	298
5.6. Effects of Agglomeration Economies on Outcomes Other Than Productivity	314
5.7. Identification of Agglomeration Mechanisms	328
5.8. Conclusion	338
Acknowledgments	340
References	341
6. Agglomeration and Innovation	349
Gerald Carlino, William R. Kerr	
6.1. Introduction	350
6.2. What is Innovation?	352
6.3. Patterns of Agglomeration and Innovation	358
6.4. Formal Theories Linking Agglomeration and Innovation	366
6.5. Additional Issues on Innovation and Agglomeration	390
6.6. Conclusions	396
Acknowledgments	397
References	398

7. Cities and the Environment	405
Matthew E. Kahn, Randall Walsh	
7.1. Introduction	406
7.2. Incorporating Local and Global Environmental Externalities into Locational Equilibrium Models	409
7.3. Global Externalities Exacerbated by the Intrametro Area Locational Choice of Households and Firms	423
7.4. Environmental Amenities in a System of Cities	427
7.5. The Urban Building Stock's Energy Consumption	445
7.6. Conclusion	457
Acknowledgment	458
References	458
 8. Urban Land Use	 467
Gilles Duranton, Diego Puga	
8.1. Introduction	468
8.2. Modeling Urban Land Use: The Monocentric Model	472
8.3. Extending the Monocentric Model	483
8.4. Agglomeration and Commercial Land Use: Modeling Polycentric Cities	503
8.5. Land Use Regulation	515
8.6. Empirical Price and Development Gradients	522
8.7. Patterns of Residential Sorting Within Cities	530
8.8. Patterns of Residential Land Development	537
8.9. Employment Decentralization and Patterns of Business Location Changes Within Cities	544
8.10. Conclusion	551
Acknowledgments	553
References	553
 9. Neighborhood and Network Effects	 561
Giorgio Topa, Yves Zenou	
9.1. Introduction	562
9.2. Neighborhood Effects	566
9.3. Network Effects	578
9.4. Neighborhood and Network Effects	599
9.5. Concluding Remarks	615
Acknowledgments	617
References	617

10. Immigration and the Economy of Cities and Regions	625
Ethan Lewis, Giovanni Peri	
10.1. Introduction	626
10.2. Immigrants' Distribution and Native Exposure	632
10.3. Theoretical Framework: The Skill Cells Approach at the National and Local Level	637
10.4. Empirical Approaches to Identify Causal Effects on Local Economies	657
10.5. Estimates of Native Responses and Effects on Outcomes	661
10.6. Recent Evolutions: Employer–Employee Panel Data and Historical Data	675
10.7. Conclusions	680
References	681
<i>Index</i>	687
 Volume 5B	
Section III. Housing and Real Estate	699
11. Housing Bubbles	701
Edward L. Glaeser, Charles G. Nathanson	
11.1. Introduction	702
11.2. The Linear Asset Pricing Model and the Idiosyncrasies of Housing	705
11.3. Empirical Regularities of Housing Dynamics	715
11.4. Rationalizing the Seemingly Irrational: Search, Heterogeneity and Agency Problems in Credit Markets	722
11.5. A Menagerie of Modest Madness: Bounded Rationality and Housing Markets	732
11.6. Public Policy and Bubbles	743
11.7. Conclusion	747
Acknowledgment	748
References	748
12. Housing, Finance, and the Macroeconomy	753
Morris A. Davis, Stijn Van Nieuwerburgh	
12.1. Introduction	754
12.2. Stylized Facts	756
12.3. Housing and the Business Cycle	767
12.4. Housing over the Life Cycle and in the Portfolio	773
12.5. Housing and Asset Pricing	787
12.6. The Housing Boom and Bust and the Great Recession	792
12.7. Housing Policy	800
12.8. Conclusion	804
Acknowledgments	805
References	805

13. The Microstructure of Housing Markets: Search, Bargaining, and Brokerage	813
Lu Han, William C. Strange	
13.1. Introduction	815
13.2. One-Sided Search	819
13.3. Random Matching	825
13.4. Pre-search, Focused Search, and Segmented Search	835
13.5. Directed Search for Housing	839
13.6. Auctions	845
13.7. Real Estate Brokers: Fundamentals	850
13.8. Competition in the Residential Real Estate Brokerage Industry	855
13.9. Incentive Issues in Real Estate Brokerage	865
13.10. Conclusions	878
Acknowledgments	879
References	879
14. US Housing Policy	887
Edgar O. Olsen, Jeffrey E. Zabel	
14.1. Introduction	888
14.2. Methods and Data	890
14.3. US Low-Income Rental Housing Policy	892
14.4. US Homeownership Policy	938
14.5. Conclusion	977
References	978
15. How Mortgage Finance Affects the Urban Landscape	987
Sewin Chan, Andrew Haughwout, Joseph Tracy	
15.1. Mortgage Finance in the United States	989
15.2. How Mortgage Finance Affects the Market for Owner-Occupied Housing	997
15.3. The Distribution of Mortgage Credit	1005
15.4. Negative Equity	1022
15.5. Foreclosures	1034
15.6. Conclusion	1039
Acknowledgments	1040
References	1040
16. Change and Persistence in the Economic Status of Neighborhoods and Cities	1047
Stuart S. Rosenthal, Stephen L. Ross	
16.1. Introduction	1048
16.2. Neighborhood Economic Status	1054

16.3. City Dynamics	1088
16.4. Conclusions and Future Research	1106
Appendix Supplemental Figures	1108
Acknowledgments	1114
References	1114

Section IV. Applied Urban Economics 1121

17. Taxes in Cities 1123

Marius Brühlhart, Sam Bucovetsky, Kurt Schmidheiny

17.1. Introduction	1124
17.2. Institutional Background	1126
17.3. Tax Setting Across Asymmetric Jurisdictions	1145
17.4. Taxation and Urban Population Sorting	1161
17.5. Taxation and Agglomeration Economies	1171
17.6. Concluding Remarks	1178
Appendix	1179
Acknowledgments	1191
References	1191

18. Place-Based Policies 1197

David Neumark, Helen Simpson

18.1. Introduction	1198
18.2. Theoretical Basis for Place-Based Policies	1206
18.3. Evidence on Theoretical Motivations and Behavioral Hypotheses Underlying Place-Based Policies	1215
18.4. Identifying the Effects of Place-Based Policies	1221
18.5. Evidence on Impacts of Policy Interventions	1230
18.6. Unanswered Questions and Research Challenges	1279
Acknowledgments	1282
References	1282

19. Regulation and Housing Supply 1289

Joseph Gyourko, Raven Molloy

19.1. Introduction	1290
19.2. Data: Old and New	1294
19.3. Determinants of Regulation	1304
19.4. Effects of Regulation	1316
19.5. Welfare Implications of Regulation	1327
19.6. Conclusion	1330
Acknowledgments	1333
References	1333

20. Transportation Costs and the Spatial Organization of Economic Activity	1339
Stephen J. Redding, Matthew A. Turner	
20.1. Introduction	1340
20.2. Stylized Facts About Transportation	1343
20.3. Theoretical Framework	1355
20.4. Reduced-Form Econometric Framework	1366
20.5. Reduced-Form Empirical Results	1372
20.6. Discussion	1383
20.7. Conclusion	1393
Acknowledgments	1394
References	1394
 21. Cities in Developing Countries: Fueled by Rural–Urban Migration, Lacking in Tenure Security, and Short of Affordable Housing	 1399
Jan K. Brueckner, Somik V. Lall	
21.1. Introduction	1400
21.2. The Empirical Aspects of Rural–Urban Migration	1402
21.3. Models of Migration and City Sizes in Developing Countries	1409
21.4. Tenure Insecurity: A Hallmark of Housing Markets in Developing Countries	1422
21.5. Provision of Affordable Housing in Developing Countries	1439
21.6. Conclusion	1448
Appendix	1450
Acknowledgments	1451
References	1451
 22. The Geography of Development Within Countries	 1457
Klaus Desmet, J. Vernon Henderson	
22.1. Introduction	1458
22.2. Development and the Aggregate Spatial Distribution	1459
22.3. Development, Space, and Industries	1475
22.4. The Urban Sector	1482
22.5. Concluding Remarks	1512
References	1513
 23. Urban Crime	 1519
Brendan O’Flaherty, Rajiv Sethi	
23.1. Introduction	1521
23.2. Criminogenic Characteristics	1522
23.3. Incentives and Deterrence	1536
23.4. Interactions	1552

23.5. Incarceration	1567
23.6. Big Swings in Crime	1583
23.7. Where are Crimes Committed?	1604
23.8. Conclusions	1612
Acknowledgments	1613
References	1613

<i>Index</i>	<i>1623</i>

FOREWORD

The fields of *Regional and Urban Economics* have evolved remarkably since 2004 when the last volume of the Handbook series (Volume 4) was published.

The emphasis of Volume 4 was very much on agglomeration at various spatial scales (neighborhood, urban, and regional). Much of the content was theoretical, with a large proportion of theoretical chapters and a clear separation between theory and empirics. Volume 4 also arrived as Krugman's New Economic Geography had reached its peak. This emphasis on agglomeration meant that many traditional urban issues were not covered. As such, policy discussions were limited to agglomeration issues, such as regional inequalities and the effect of market integration (following worries associated with "globalization" and deeper economic integration within Europe and North America). The decade since Volume 4 has seen continued progress on agglomeration and related areas, but it has also seen a significant broadening in both the areas of study and the methods of inquiry.

This volume is in part a return to more traditional urban topics that were covered in Volumes 1–3 of the Handbook series. One example of this is housing, a research topic which has seen major advances in the last 10 years. A major housing crisis in the United States and much of the developed world is certainly part of the explanation for revival of research on housing. In addition, there are also important ongoing debates about urban sprawl and its effects and how land use regulations are shaping cities in the United States and elsewhere. Technology and sometimes legislation are also changing the way we buy and sell houses. This raises some interesting questions about the microstructure of the housing market. Thus, Volume 5 of the *Handbook of Regional and Urban Economics* has a significant emphasis on housing and property markets.

Housing is not the only new focus for urban research. There is also renewed interest in the effects of transportation on cities, neighborhood and city dynamics, urban amenities, urban environmental issues, urban crime, urban costs, land use, migration, and a range of other topics. These issues are considered in both developed and developing world settings. Volume 5 reflects this intellectual broadening as well.

Another important shift in urban and regional economics is in methods. For the first time in the *Handbook of Regional and Urban Economics* series, explicit chapters on methodology are included. The greater availability of data and the gradual adaption of "modern" methodologies have profoundly changed the nature of empirical work. These approaches (structural and quasi-experimental) are becoming more widely adopted. The chapters in this volume acknowledge this, but they also point out that a lot urban and regional research remains in need of a methodological upgrade. In addition, the chapters point to a range of unique methodological challenges arising from the spatial data that is

used in urban and regional research. The direct application of methodologies borrowed from labor economics or industrial organization is, thus, often not enough. Fortunately, both the chapters focusing primarily on methods and those that consider individual topics offer numerous suggestions of how to move forward. In most instances, this involves forging closer links between theory and empirical research.

All of these issues have significant implications for public policy. Volume 5 includes chapters focusing on policy topics that have had little coverage in previous volumes, such as mortgages, place-based policies, and urban crime. The volume also includes chapters on more traditional issues such as tax competition, neighborhood effects, and housing policy. These traditional issues are still extremely important but are now explored using more credible empirical approaches. And although these chapters are particularly oriented toward policy, the applied nature of Urban and Regional Economics means that most chapters are policy relevant at least to some degree.

Ultimately, we see the chapters included in the volume as making a strong case for research that appropriately combines theory and empirics, that embraces the many elements of urban economies, and that is policy relevant. Of course, as the volume has come together, it has become apparent that there are gaps in the volume just as there are gaps in the fields of regional and urban economics. For instance, too much of the empirical evidence on urban issues comes from American cities. While the volume does contain two chapters focused on issues in developing countries, more work on urban phenomena in developing countries is needed. As another example, while there is a chapter on transportation focused on evaluation of major inter regional transport networks, there is no coverage of traditional and evolving topics such as modal choice, peak pricing, the use of incurred transport costs to value urban amenities, and the like. We hope that these and other gaps will motivate young (and less young) researchers to expand our knowledge.

We are grateful to many people and organizations for helping to make this project happen. The contribution of the authors is obvious. These contributions were sharpened by the participants at conferences sponsored by the Wharton Real Estate Department and the Centre for Real Estate at the Rotman School of Management at the University of Toronto. Several papers were also presented at the Urban Economics Association sessions at the North American Regional Science Council meetings and at the National Meetings of the American Real Estate and Urban Economics Association. We are grateful to the people and organizations who have made these interactions possible. We also are grateful to various people at Elsevier for their helpfulness and professionalism, especially Joslyn Chaiprasert-Paguio and Scott Bentley. Finally, we are all grateful to all those who are close to us for their patience and support.

Gilles Duranton
Vernon Henderson
William Strange
November 4, 2014

CONTRIBUTORS

Nathaniel Baum-Snow

Department of Economics, Brown University, Providence, RI, USA

Kristian Behrens

Department of Economics; CIRPÉE, Université du Québec à Montréal, Montréal, QC, Canada;
National Research University, Higher School of Economics, Moscow, Russia, and CEPR,
London, UK

Marius Brühlhart

University of Lausanne, Lausanne, Switzerland, and Centre for Economic Policy Research
(CEPR), London, UK

Jan K. Brueckner

Department of Economics, University of California, Irvine, CA, USA

Sam Bucovetsky

York University, Toronto, ON, Canada

Gerald Carlino

Federal Reserve Bank of Philadelphia, Philadelphia, PA, USA

Sewin Chan

Robert F. Wagner School of Public Service, New York University, NY, USA

Pierre-Philippe Combes

Aix-Marseille University (Aix-Marseille School of Economics), CNRS & EHESS, Marseille;
Economics Department, Sciences Po, Paris, France, and Centre for Economic Policy
Research (CEPR), London, UK

Morris A. Davis

Department of Finance and Economics, Rutgers Business School, Rutgers University, Newark,
NJ, USA

Klaus Desmet

Department of Economics, Southern Methodist University, Dallas, TX, USA

Gilles Duranton

Wharton School, University of Pennsylvania, Philadelphia, PA, USA, and CEPR, London, UK

Fernando Ferreira

The Wharton School, University of Pennsylvania, Philadelphia, PA, USA

Steve Gibbons

London School of Economics, London, UK

Edward L. Glaeser

Harvard University and NBER, Cambridge, MA, USA

Laurent Gobillon

Centre for Economic Policy Research (CEPR), London, UK; Institut National d'Etudes Démographiques; Paris School of Economics, Paris, France, and The Institute for the Study of Labor (IZA), Bonn, Germany

Joseph Gyourko

The Wharton School, University of Pennsylvania, Philadelphia, PA, and NBER, Cambridge, MA, USA

Lu Han

Rotman School of Management, University of Toronto, Toronto, ON, Canada

Andrew Haughwout

Federal Reserve Bank of New York, NY, USA

J. Vernon Henderson

Department of Geography, London School of Economics, London, UK

Thomas J. Holmes

University of Minnesota and Federal Reserve Bank of Minneapolis, Minneapolis, MN, USA

Matthew E. Kahn

Department of Economics, UCLA and NBER and IZA, Los Angeles, CA, USA

William R. Kerr

Harvard University, Bank of Finland, and NBER, Boston, MA, USA

Somik V. Lall

Urban Development and Resilience Unit, Sustainable Development Network, World Bank, USA

Ethan Lewis

Dartmouth College, Hanover, NH, and NBER, Cambridge, MA, USA

Raven Molloy

Board of Governors, Federal Reserve System, Washington, DC, USA

Charles G. Nathanson

Northwestern University, Evanston, IL, USA

David Neumark

UCI, NBER, and IZA, Irvine, CA, USA

Brendan O'Flaherty

Department of Economics, Columbia University, NY, USA

Edgar O. Olsen

Department of Economics, University of Virginia, Charlottesville, VA, USA

Henry G. Overman

London School of Economics, London, UK

Eleonora Patacchini

Cornell University, Ithaca, NY, USA

Giovanni Peri

University of California-Davis, CA, and NBER, Cambridge, MA, USA

Diego Puga

CEPR, London, UK, and Centro de Estudios Monetarios y Financieros (CEMFI), Madrid, Spain

Stephen J. Redding

Economics Department and WWS, Princeton University Fisher Hall, Princeton, NJ, USA

Frédéric Robert-Nicoud

CEPR; SERC, The London School of Economics and Political Science, London, UK, and Geneva School of Economics and Management, Université de Genève, Genève, Switzerland

Stuart S. Rosenthal

Maxwell Advisory Board Professor of Economics, Department of Economics, Syracuse University, Syracuse, NY, USA

Stephen L. Ross

Department of Economics, University of Connecticut, Storrs, CT, USA

Kurt Schmidheiny

Centre for Economic Policy Research (CEPR), London, UK; University of Basel, Basel, Switzerland, and CESifo, Munich, Germany

Rajiv Sethi

Department of Economics, Barnard College, Columbia University, NY, USA, and Santa Fe Institute, Santa Fe, NM, USA

Holger Sieg

University of Pennsylvania, Philadelphia, PA, USA

Helen Simpson

University of Bristol, CMPO, OUCBT and CEPR, Bristol, UK

William C. Strange

Rotman School of Management, University of Toronto, Toronto, ON, Canada

Giorgio Topa

Federal Reserve Bank of New York and IZA, NY, USA

Joseph Tracy

Federal Reserve Bank of New York, NY, USA

Matthew A. Turner

Economics Department, Brown University, Providence, RI, USA

Stijn Van Nieuwerburgh

Department of Finance, Stern School of Business, New York University, NY, USA

Randall Walsh

Department of Economics, University of Pittsburgh and NBER, Pittsburgh, PA, USA

Jeffrey E. Zabel

Department of Economics, Tufts University, Medford, MA, USA

Yves Zenou

Stockholm University, IFN, and CEPR, Stockholm, Sweden

This page intentionally left blank

SECTION III

Housing and Real Estate

This page intentionally left blank

CHAPTER 11

Housing Bubbles

Edward L. Glaeser*, **Charles G. Nathanson†**

*Harvard University and NBER, Cambridge, MA, USA

†Northwestern University, Evanston, IL, USA

Contents

11.1. Introduction	702
11.2. The Linear Asset Pricing Model and the Idiosyncrasies of Housing	705
11.2.1 The linear asset pricing model or user cost model	706
11.2.2 Processes with momentum, mean reversion and stochastic growth rates	707
11.2.3 What is special about housing?	709
11.2.4 Directions for future research	714
11.3. Empirical Regularities of Housing Dynamics	715
11.3.1 Excess variance	715
11.3.2 The shape of the cycle: short run momentum and long run mean reversion	716
11.3.3 Quantities versus prices	718
11.3.4 Bubbles in history	720
11.3.5 Directions for future research	721
11.4. Rationalizing the Seemingly Irrational: Search, Heterogeneity and Agency	
Problems in Credit Markets	722
11.4.1 Search, learning, and momentum	722
11.4.2 Changes in credit conditions	724
11.4.3 Agency, underpriced default options and rational bubbles	728
11.4.4 Directions for future research	732
11.5. A Menagerie of Modest Madness: Bounded Rationality and Housing Markets	732
11.5.1 Exogenous and heterogeneous beliefs: search, endogenous housing supply, and land acquisition	733
11.5.2 Extrapolative beliefs	734
11.5.3 Cognitive limitations: natural expectations, spatial benchmarking, and rule-of-thumb spending	737
11.5.4 Social learning and the entrepreneurs of error	741
11.5.5 Directions for future research	743
11.6. Public Policy and Bubbles	743
11.7. Conclusion	747
Acknowledgment	748
References	748

Abstract

Housing markets experience substantial price volatility, short-term price change momentum, and mean reversion of prices over the long run. Together, these features, particularly at their most extreme, produce the classic shape of an asset bubble. In this chapter, we review the stylized facts of housing bubbles and discuss theories that can potentially explain events like the boom–bust cycles of the 2000s. One set of theories assumes rationality and uses idiosyncratic features of the housing market, such as intensive search and short-selling constraints, to explain the stylized facts. Cheap credit provides a particularly common rationalization for price booms, but temporary periods of low interest rates will not explain massive price swings in simple rational models. An incorrectly underpriced default option can make rational bubbles more likely. Many nonrational explanations for real estate bubbles exist, but the most promising theories emphasize some form of trend chasing, which in turn reflects boundedly rational learning.

Keywords

Real estate, Housing, Behavioral economics, Land, Bubble, Speculation

JEL Classification Code

R21

11.1. INTRODUCTION

Between 2000 and 2012, America experienced a great housing convulsion that had all the classic features associated with real estate bubbles. Housing prices rose dramatically and then fell, leaving average real housing prices in 2012 no higher than they were in 2000. Price growth, between 2000 and 2006, was much higher in some places than in others, and the places with the biggest price growth experienced the largest declines. Surprisingly, some of the biggest booms occurred in places like Phoenix and Las Vegas, which appear to have few short-run limits on new construction (Davidoff, 2013; Nathanson and Zwick, 2013; Gao, 2014).

During the years of biggest boom—2003, 2004, and 2005—when the change in real housing price growth is regressed on the 1-year lag of price growth across metropolitan areas, the coefficient is greater than one. Price growth seemed to build upon itself. This phenomenon represents the more general tendency of price growth to show strong positive serial correlation at 1-year frequencies (Case and Shiller, 1989). There was also a clear pattern of spatial correlation, where a boom that started on the coast seems to have spread to neighboring inland metropolitan areas (Ferreira and Gyourko, 2012).

The US housing cycle that occurred between 2000 and 2012 is extreme but hardly unique. Other countries, such as Ireland and Spain, also experienced housing booms and crashes over those years. While Japan's housing market remained stable after 2000, Japan had experienced its own massive real estate cycle in the 1980s and early 1990s. American history is replete with examples of real estate booms and busts, from the days of the early

Republic to the American convulsion of the early 1980s. In summarizing these events, Glaeser (2013) argued that while these events may clearly look like bubbles *ex post*, even at their height, prices could be reconciled with standard models of real estate evaluation.

In Section 11.2 of this essay, we present a benchmark, rational model so that we can assess departures from this model when we discuss stylized facts. We use a simple user cost model of housing value that we refer to as the linear asset pricing model or LAPM, following Head et al. (2014). This approach runs deep within real estate and housing economics, but it differs from the general equilibrium approach preferred by macroeconomists and discussed by Davis and Van Nieuwerburgh in Chapter 12. In any user cost model that descends from Poterba (1984), prices equal the expected value of the exogenous flow of discounted future benefits from homeownership. While some empirical facts may be obviously anomalous, many housing facts are only surprising in the context of the model, which is why we begin with a simple model, rather than stylized facts.

The LAPM approach not only is rational but also abstracts away from critical institutional features of the housing such as search and heterogeneity. Section 11.2 also discusses important institutional features of the housing market that differ from other assets including the extreme difficulty of short sales, the extreme heterogeneity of the asset, the dominant role of amateur investors, and the limited information about current asset values. These differences do limit the widespread applicability of the user cost or LAPM formula, but it remains a useful benchmark with which to examine housing price fluctuations.

Section 11.3 discusses stylized facts about housing markets. Housing markets do experience excess variance of price changes relative to the LAPM's predictions, but excess variance is less endemic than in other asset markets and shows up primarily in brief outbursts. A few overoptimistic traders will more readily dominate standard asset markets than the highly democratic housing market, which might explain the apparently more stable nature of housing markets most of the time (Nathanson and Zwick, 2013).

Section 11.3 also notes the spatial heterogeneity of the recent boom and bust, the strong short-term positive serial correlation of prices, and the even stronger long-term negative price change serial correlation. The short-term positive serial correlation is difficult to square with the LAPM, but negative serial correlation over longer time horizons is more readily reconcilable with rationality if housing supply responses are delayed. We also discuss the movements in the quantities of housing, which can refer to both the volume of housing produced and the volume of housing sold. We end this section with a discussion of real estate bubbles in US history that draws heavily from Glaeser (2013).

Section 11.4 turns to rationalizing the seemingly irrational: economic approaches to understanding these gigantic fluctuations, which maintain individual rationality assumptions. Models of search by heterogeneous consumers, such as Novy-Marx (2009), Genesove and Han (2012), Head et al. (2014), and Guren (2014), can amplify shocks to the housing demand and thereby generate momentum in prices over shorter time

periods. Heterogeneous demand can also generate price volatility, if there are exogenous shocks to supply. We discuss the possible role of rational learning but conclude that fully Bayesian updating has little power on its own to explain the stylized facts.

Our final attempt to explain the events with rational models is to discuss the possible role of credit. [Mian and Sufi \(2014\)](#) have compellingly made the case that subprime lending did push up prices in many areas. While lower interest rates will have only a modest impact on prices in the standard LAPM, assuming that individuals anticipate the mean reversion of interest rates ([Cox et al., 1985](#)), an underpriced default option will allow rational bubbles to occur more readily. One natural explanation for charging borrowers too little for the risk of default is the existence of agency problems either within financial institutions or between mortgage originators and the eventual owners of securitized mortgages. We do not formally model the financial institution's structure but assume that for some rational reason, the agents of lenders are pushing cash out the door too quickly and cheaply. Naturally, overly cheap credit might also reflect overoptimism on the part of lending institutions or some other form of limited rationality.

Typically, rational bubbles in housing require a violation of the standard transversality condition ([Malinvaud, 1953](#)): with such bubbles, the discounted infinite horizon expected value of the home is strictly positive.¹ Violations of the transversality condition make the problems of builders and sellers difficult to explain: why ever sell a house today when its price is rising so quickly? As such, rational bubbles are particularly difficult to deliver with even moderately elastic supply ([Glaeser et al., 2008](#)). If default risk is underpriced, then rational bubbles can occur even with moderately elastic supply and without violating any transversality condition. Naturally, underpriced default risk might itself reflect bounded rationality, or alternatively, it might reflect an agency problem within lending institutions.²

In [Section 11.5](#), we present a collection of less rational models of housing bubbles. We begin with models in which beliefs are exogenous including [Glaeser et al. \(2008\)](#), [Piazzesi and Schneider \(2009\)](#), and [Nathanson and Zwick \(2013\)](#). These models are useful for exploring the ways in which institutional features of the housing market interact with irrationality. For example, [Glaeser et al. \(2008\)](#) and [Nathanson and Zwick \(2013\)](#) both considered land availability; [Piazzesi and Schneider \(2009\)](#) considered search dynamics. While this line of work is potentially useful in generating ancillary predictions about housing bubbles (e.g., they are most common in areas where land supply is limited in the long run), they cannot explain the ultimate source of beliefs about house prices.

¹ The transversality condition helps ensure that a dynamic competitive equilibrium is Pareto optimal. This does not imply that the transversality actually holds in actual asset markets, but its critical role in a number of important theorems should make us wary of dropping it arbitrarily.

² For example, if banks can't observe credit quality but can observe loan amounts and if lending takes effort, then the bank may want to reward loan officers based on the number of loans made to induce more effort. This may lead to low-quality loans.

We then turn to a second major class of models: extrapolation. In one class of extrapolative models, the belief formation process is ad hoc but assumed for a small class of traders. In a second class, the belief formation process follows from imperfect learning about a stochastic growth rate. In a sense, the first class assumes massive irrationality for the few, while the second class assumes modest irrationality for the many. It may be that the first is more appropriate in standard financial markets where a small number of well-financed traders can drive markets, while the second class is more appropriate in housing markets.

A third class of models discusses other forms of limited rationality, including limited memory, basing models on short runs of data, rule-of-thumb buying strategies, and price estimates based on natural geographic comparisons. This discussion highlights the fact that once perfect rationality is dropped, an essentially infinite array of assumptions is possible. This is Tolstoy's corollary: There is only one correct answer to an optimization problem, but there are an uncountable number of wrong answers.

Yet, there are some insights from near-rational that seem more important than others. [Fuster et al. \(2010\)](#) introduced natural expectations where individuals forecast the future using simple models, perhaps incorporating only a limited amount of history. Short-history forecasting may lead homebuyers to ignore low-frequency mean reversion and the longer-term impact of elastic supply, which will eventually reduce prices. [Glaeser \(2013\)](#) argued that the failure to internalize the long-run impacts of elastic supply on price is a constant feature of US real estate history. We end [Section 11.5](#) by discussing social learning and the possible role of entrepreneurs, like real estate brokers, who attempt to persuade buyers that housing prices can only go up.

Finally, in [Section 11.6](#), we discuss public policy implications of real estate bubbles. Two facts seem clear: Real estate fluctuations exist, and they have displayed a remarkable ability to wreak havoc on financial systems. While it may be impossible or infeasible to prick bubbles while they expand, it may still be possible to undertake protective actions to ensure that the bursting of the bubble will not cause as much damage in the future. There may also be social benefits from reconsidering those policies that encourage individuals to borrow heavily to invest in real estate. [Section 11.7](#) concludes.

11.2. THE LINEAR ASSET PRICING MODEL AND THE IDIOSYNCRASIES OF HOUSING

In [Section 11.4](#), we will present the core stylized facts about the housing market that need to be addressed by the models of housing bubbles. For these facts to help inform a survey of housing bubbles, we must start with some clarifying discussion and algebra. Bubbles are typically defined as periods in which asset prices “run well above or below the intrinsic value” ([Fama, 1965](#)). In the real estate context, fundamentals can mean either the value based on the flow of rents, as in the “user cost” model and its LAPM variant, or the flow of well-being associated with living in a particular spot.

Housing markets are different from other asset markets. There exist real estate linked assets, including real estate investment trusts (REITs) or collateralized mortgage securities, that trade in large markets, but, more typically, housing is bought and sold in small, decentralized transactions. This fact is obviously true in the enormous market for single-family housing, but it is also true in the realm of commercial real estate as well. The dollar amounts of each sale may be larger, but ultimately, the purchase of Rockefeller Center is just as idiosyncratic as the purchase of that Tudor home on the corner, if not more so.

11.2.1 The linear asset pricing model or user cost model

Nonetheless, the benchmark model of housing prices, often called the “user cost” model, assumes that there is a single price of housing that is paid by all buyers, who are typically homeowners.³ This price follows from an intertemporal no-arbitrage condition. The value of owning a home equals the benefits today plus the asset value tomorrow, or the price P_t equals $R_t + \frac{E(P_{t+1})}{1+r}$, where R_t represents net benefits of owning during time t , and $\frac{1}{1+r}$ represents a constant discount factor. The “fundamental” value then equals $E\left(\sum_{j=0}^{\infty} (1+r)^{-j} R_{t+j}\right)$. Empirically, the values of R_{t+j} are either associated with observed market rents (Himmelberg et al., 2005) or the benefits of living in a particular area including amenities and income (Glaeser et al., 2014; Head et al., 2014). These raw series can provide testable implications for the observed series of prices.

One important aspect of this approach is that it avoids any consideration of risk aversion and portfolio composition. These issues are particularly important to the financial economists and macroeconomists who study housing, and they are addressed in Chapter 12 by Davis and Van Nieuwerburgh.

Shiller (1981) provided a famous, nonparametric approach to testing for excess variance by noting that since $\sum_{j=0}^{\infty} (1+r)^{-j} R_{t+j}$ equals $E\left(\sum_{j=0}^{\infty} (1+r)^{-j} R_{t+j}\right) + \text{Error}$ and since the error should be independent of the expectation, the variance of the fundamental must be greater than the variance of the price series. Of course, this is not true in the US stock market, and it is not true in housing markets either. This calculation is also somewhat compromised by time-varying discount factors, which we will address later.

For work on housing markets, it is somewhat more common to assume a particular stochastic process for the fundamental, R_t , and to work with the formula implied by that quantity. We briefly consider the implications of four stochastic processes: $R_t = (1+g_R)R_{t-1} + \epsilon_t^R$ (constant growth, i.i.d. error), $R_t = R_{t-1} + \epsilon_t^D + \theta\epsilon_{t-1}^D$ (no growth, moving average error), $R_t = \delta R_{t-1} + (1-\delta)\bar{R} + \epsilon_t^D + \theta\epsilon_{t-1}^D$ (mean reverting, moving average error, e.g., ARMA(1,1)), and $R_t = R_{t-1} + g_t + \epsilon_t^D$, where $g_t = \lambda g_{t-1} + (1-\lambda)\bar{g} + \epsilon_t^g$ (stochastic growth rate). The moving average process can create

³ Haughwout et al. (2011) reminded us that during the boom, many of the buyers of single-family homes were actually investors.

some persistence in price changes but much less than the persistence in rent changes. The mean-reverting process can create robust amounts of price mean reversion but very small amounts of price momentum. The stochastic growth rate process can predict both mean reversion and at least modest momentum. That process can also yield relatively high levels of price variance.

The simplest, nonstochastic process is to assume a constant growth rate for R_t , so that $R_t = (1 + g_R)R_{t-1} + \epsilon_t^R$, where ϵ_t^R is an independently and idiosyncratically distributed (i.i.d.) noise term. In that case, the LAPM price becomes $\frac{(1+r)R_t}{r-g_R}$. One implication of this formula is that small differences in g_R can generate extreme differences in prices, especially when interest rates are low, as [Himmelberg et al. \(2005\)](#) emphasized. Since the LAPM version of the user cost model can imply extremely high price-rent ratios with seemingly reasonable parameter estimates, it can also make bubbles seem rational. The use of the LAPM in 2006 to justify the high prices at the peak of the cycle suggests just how difficult it can be to identify a bubble while it is happening.

[Shiller's \(1981\)](#) insight was that even if it is hard to tell which price levels are compatible with an LAPM price, the model also yields predictions about price change variances and covariances that can be rejected by the data. The constant growth rate with i.i.d. error assumption implies that the standard deviation of price changes equals $\frac{1+r}{r-g_R}$, or the price-rent ratio, times the standard deviation of ϵ_t^R , which is also the standard deviation of changes in the rent. In other words, the ratio of standard deviation of price changes to the standard deviation of rent changes equals the price-rent ratio. That fact holds true for all time intervals, because the shocks are i.i.d. Moreover, this simple random walk with drift predicts neither positive serial correlation in short-term price changes (momentum) nor negative serial correlation in long-term price changes (mean reversion).

11.2.2 Processes with momentum, mean reversion and stochastic growth rates

For the next two illustrative processes, we assume that $g_R = 0$. Given this simplification, we next complicate the shocks with a moving average component and assume that $R_t = R_{t-1} + \epsilon_t^D + \theta\epsilon_{t-1}^D$, where ϵ_t^D is i.i.d. and $\theta \leq 1$. The implied LAPM price is $\frac{(1+r)R_t + \theta\epsilon_t^D}{r}$. The ratio of the standard deviation of price changes to the standard deviation of rent changes is $\frac{1}{r} \sqrt{\frac{(1+r+\theta)^2 + \theta^2 r^2}{1+\theta^2}}$. The serial correlation of rent changes is $\frac{\theta}{1+\theta^2}$, and the implied serial correlation in price changes is $\frac{(1+r+\theta)r\theta}{(1+r+\theta)^2 + \theta^2 r^2}$, which must be less than $\frac{r\theta}{1+\theta^2}$. This implies that the serial correlation of prices must be less the interest rate times the serial correlation in prices. If $r = 0.1$, then an increase in the value of θ from 0 to 1 causes the serial correlation of rents to increase from 0 to 0.5, but the serial correlation of housing prices will rise only from 0 to 0.0475. Very large amounts of serial correlation in rents are associated with quite modest amounts of serial correlation in price changes.

This moving average process does little to generate mean reversion at lower frequencies. To allow for this possibility, we assume that $R_t = \delta R_{t-1} + (1 - \delta)\bar{R} + \epsilon_t^D + \theta \epsilon_{t-1}^D$. The LAPM price satisfies

$$P_t = \frac{(1+r)R_t}{1+r-\delta} + \frac{(1+R)(1-\delta)\bar{R}}{r(1+r-\delta)} + \frac{\theta \epsilon_t^D}{1+r-\delta} \quad (11.1)$$

Generating mean reversion at longer frequencies is feasible with this type of autoregressive process, but generating substantial momentum is even more difficult than in the simpler moving average process discussed above. A shock from last period can increase price growth today, as long as $\theta > \frac{(1+r)(1-\delta)}{r}$. This will hold if $\delta = 1$, which is the random walk case discussed above, but with higher levels of mean reversion, even one-period momentum becomes less and less plausible, and it is still impossible to get price change momentum that is nearly as big as rent change momentum.

Last, we consider stochastic growth rates. We assume that $R_t = R_{t-1} + g_t + \epsilon_t^D$, where $g_t = \lambda g_{t-1} + (1 - \lambda)\bar{g} + \epsilon_t^g$. The growth rate is persistent in the short term but ultimately reverts to the area level norm. These assumptions also imply that the LAPM price satisfies

$$P_t = \frac{(1+r)R_t}{r} + \frac{(1+r)^2(1-\lambda)\bar{g}}{r^2(1+r-\lambda)} + \frac{(1+r)\lambda g_t}{r(1+r-\lambda)} \quad (11.1')$$

The ratio of the standard deviation of price changes to the standard deviation of rent changes is again roughly similar to price-rent ratios. The one-period autocorrelation of rental shocks is $\frac{\lambda}{\varphi(1-\lambda^2)+1}$, where $\varphi = \frac{\text{Var}(\epsilon_t^D)}{\text{Var}(\epsilon_t^g)}$. The one-period autocovariance of price

changes equals $\frac{(1-\lambda^2+r)(1+r)^2\lambda}{r(1+r-\lambda)^2(1-\lambda^2)}\text{Var}(\epsilon_t^g)$, which is distinctly positive. The long stickiness of growth rates in fundamentals provides the best chance of hitting high degrees of serial correlation in housing prices.

We have required the LAPM, so far, to only concern fundamentals, but as with most asset pricing equations, the one-period indifference condition $R_t + \frac{1}{1+r}E(P_{t+1}) = P_t$ admits “rational bubbles,” where $R_t + \frac{1}{1+r}E(P_{t+1} + \text{Bubble}_{t+1}) = P_t + \text{Bubble}_t$. The key requirement is that $\frac{1}{1+r}E(\text{Bubble}_{t+1}) = \text{Bubble}_t$. Such a bubble would violate the standard transversality condition requiring the discounted value of future housing prices to converge to zero as t goes to infinity. The bubble’s discounted value at all time periods will always equal its value today.

In their most plausible formulation, rational bubbles are stochastic, so, for example, the bubble will burst in each period with probability ν . In this case, the bubble multiplies by $\frac{1+r}{1-\nu}$ during each period that it doesn’t burst. While this formulation is mathematically conceivable, it requires buyers to expect that with some probability, housing will become extraordinarily expensive within some reasonable probability. For example, a Las Vegas house in 2006 was valued at \$300,000 and one-half of that was a rational bubble: the

market dropped by over 50% after the bust. Moreover, assume that buyers thought that there was a 50% chance that the bubble would burst in each year and that $r=0.05$. If the bubble didn't burst, the home would have been worth \$465,000 in 2007 (with probably 0.5), \$811,500 in 2008 (with probably 0.25), and \$1.54 million in 2009 (with probably 0.125).

Is such price growth remotely plausible and is it compatible with other features of housing markets? Glaeser et al. (2008) argued that such price growth is incompatible with even moderately elastic supply. Presumably, builders in Las Vegas in 2009 would do anything to build houses that can be sold for such high prices. Moreover, this \$1.54 million dollar house would have to find buyers, at least some of whom need down payments. Would it be possible for price differences between similar regions to widen so extraordinarily and persist in expectation? As we will discuss later, we find this rational bubble formulation far more plausible if it occurs in a setting where lenders are providing borrowers with an underpriced default option.

11.2.3 What is special about housing?

While the LAPM model treats housing essentially as a standard security, this assumption is badly at odds with reality. Housing is idiosyncratic and traded individually. Searching for housing can be a lengthy process. We now discuss how the salient aspects of the housing market make housing somewhat different from other securities.

This dispersed, idiosyncratic market means that there is no such thing as the current price of housing, in the same way as there is a current price of General Electric stock. Moreover, across the United States and even across the larger metropolitan areas, the heterogeneity of markets is enormous. Between 2004 and 2006, FHFA price data show the value of homes in Phoenix increasing by over 50%. In nearby Albuquerque, New Mexico, prices rose only 17% over the same time period. In Fort Collins, Colorado, real prices actually fell during those 2 years.

Abel and Deitz (2010) divided American housing markets into four groups based on their experiences from 2000 to 2008. Many had booms and busts, and many avoided both booms and busts. A smaller number experienced booms without busts, like Casper, Wyoming. Even fewer areas, most notably Detroit, Michigan, experienced busts without booms. Cyclical activity also differs substantially across countries, and ideally, housing models could also explain this heterogeneity.⁴

The fragmented housing market may help explain why post-boom drops are so much slower in housing markets than in other asset classes. The history of the US stock market is punctuated by particular days of infamy during which shares tumbled, such as 24 October 1929 (Black Tuesday), 19 October 1987 (Black Monday), and 15 October 2008. There

⁴ Differential housing supply elasticities across countries may explain some of these differences (Caldera and Johansson, 2013).

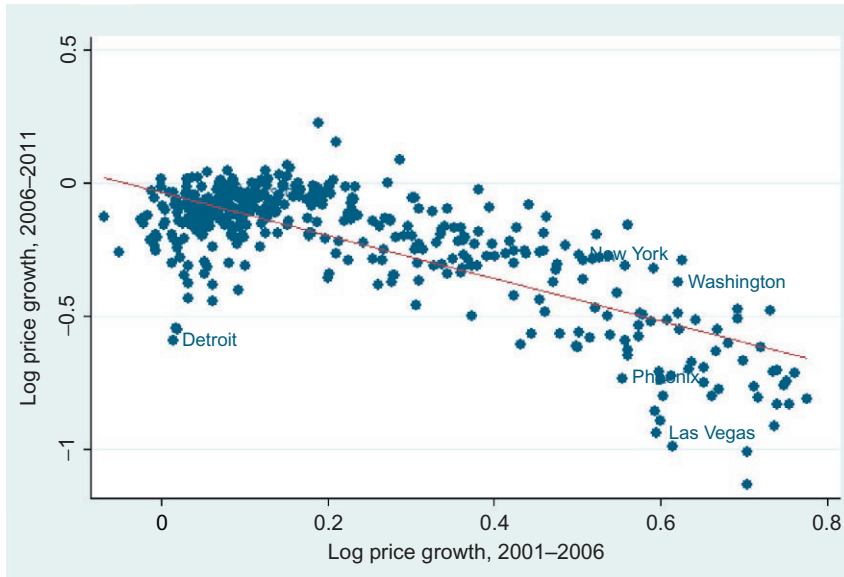


Figure 11.1 Real price growth (FHFA), 2001–2006 versus 2006–2011.

is no comparable date in which housing prices plummet. During the last two general housing market slumps, it took several years for prices to hit the bottom.

This slow adjustment process may also reflect transaction costs in housing, which slow sales and even price adjustments or the markets' domination by "amateurs." Ordinary homebuyers and sellers may have less up-to-date information than professional stock traders. These amateurs may also suffer more from "loss aversion" (as in [Genesove and Mayer, 2001](#)), which makes sellers unwilling to cut their asking prices even in the face of a market collapse.⁵ This loss aversion might also explain why market volumes drop dramatically during busts. [Scheinkman and Xiong \(2003\)](#) connected the high levels of trading during financial booms to heterogeneous beliefs, and that also seems likely to be relevant in housing markets.⁶

The market fragmentation of housing also helps explain why housing market "facts" tend to be cross-sectional as much as time series. For example, the mean reversion of housing prices is often illustrated by a cross-sectional graph, such as [Figures 11.1 and 11.2](#), showing the strong correlation between the degree of price growth during a boom period and the degree of price decline during the bust. Indeed, the tendencies of local real estate markets towards excess variance, high-frequency momentum, and lower-frequency mean reversion show up even controlling for national market trends.

⁵ We thank William Strange for emphasizing these points.

⁶ [Hong et al. \(2006\)](#) also connected bubbles and volume, but their analysis also relies on lock-up constraints, such as those that faced Internet entrepreneurs.

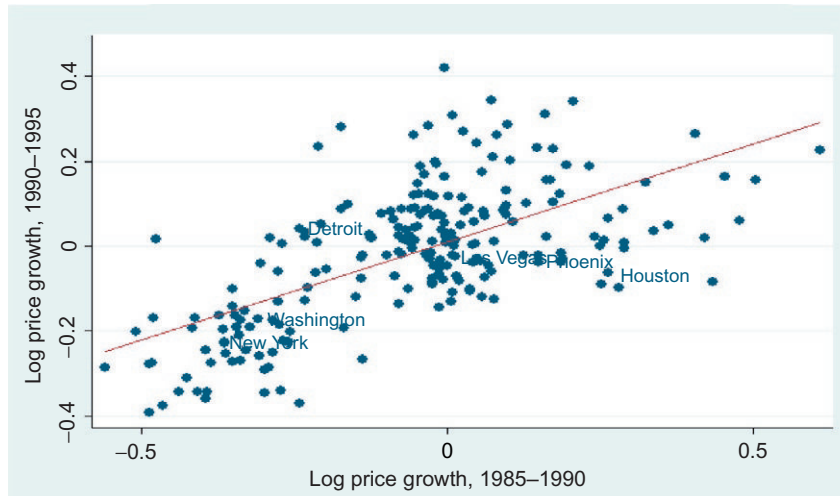


Figure 11.2 Real price growth (FHFA), 1990–1995 versus 1985–1990.

The fragmented nature of local housing markets also explains why housing economists rely on two distinct no-arbitrage conditions to measure “appropriate” housing values. In the temporal user cost model, the flow benefit of owning a house at time t , plus the discounted value of the home at time $t + 1$, must equal the cost of buying the house at time t (P_t) plus other costs such as taxes and effort involved in maintenance. The alternative, spatial no-arbitrage condition is that the cost of living in Place A must equal the cost of living in Place B plus whatever extra benefits accrue to living in Place A relative to Place B, as in the Alonso–Muth–Mills model. Both equations are complicated by idiosyncrasies of mortgage borrowing, risk aversion, and transaction costs, but they provide useful starting points for thinking about asset prices and real estate bubbles.

Implementing this simple intertemporal user cost formula is challenged by the difficulty of observing most of the key parameters. Prices may be seen by the econometrician but little else can be directly observed. While the benefit of owning stocks for ordinary stockholders yields only dividend yields and price appreciation, the primary benefit of owning a home is that you get to live in the home and that may yield different benefits for different people. Moreover, ownership carries other costs. Some of these costs are directly observable, like property taxes, while others, such as the sweat of home care, are not. Limited observability of costs and benefits means that the no-arbitrage condition in housing will always be far less precise than the equivalent conditions in other asset markets.

The most straightforward means of quantifying the benefits of owning is to use rents. If owning and renting were otherwise identical, then the benefits of owning should be equal to the benefits of renting. Yet, there are three reasons why identifying the flow

value of owning housing with the prevalent rental rates is problematic: The homes aren't the same, the neighborhoods aren't the same, and the unobserved costs aren't the same. Glaeser and Gyourko (2009) documented that the observable differences between rental and owner-occupied structures are enormous: 64% of owner-occupied units are single-family detached as opposed to 18% of rental units. Owner-occupied units are also less likely to be located in central cities and more likely to be in neighborhoods that are rated as excellent by their residents. Finally, Goodman (2005) used the 2003 American Time Use Survey and documented the significantly large amounts of time spent on home and yard maintenance by owners, but it is hard to quantify the costs of that effort.

An alternative approach is to eschew rental data as being nonrepresentative and instead focus on measuring the benefits of locating in one metropolitan area, rather than another, following Rosen (1979) and Roback (1982). Local earnings and amenities are the typical sources of local benefits, but amenities are difficult to completely capture and the earnings of the marginal homebuyer are also not directly observed. Finally, since this approach is inherently comparative, it can only answer whether the fluctuations in price in one area seem reasonable relative to the fluctuations in price in some other place.

Housing is also different from stocks and bonds because housing is the democratic asset, owned by over 60% of American households. Policies and preferences come together to ensure that homeownership is dispersed among millions of Americans, rather than concentrated in the hands of professional investors who rent them out. For these households, especially those with long horizons, housing looks more like a consumption good than a financial asset. Some caution therefore is needed while applying models typically used with the stock market to the housing market, especially since we are all born short housing. Of course, some parts of the real estate market—like commercial properties, undeveloped land, and rental residential houses—are owned by investors calculating discounted future cash flows, and residential housing resembles a financial asset more in areas with these other types of properties (Nathanson and Zwick, 2013).

The widespread nature of housing ownership also creates policy-related issues relative to housing booms that are less present with other asset classes. Historically, housing risk was held by homeowners and by savings institutions. When a boom crashed, millions of ordinary people were worse off, and the banking system was imperiled. The widespread nature of the pain and the potentially disastrous consequences of a banking system meltdown make serious policy response far more likely than when there is a bust in some smaller asset class. Few policy makers argued that indebted investors in Internet stocks had no obligation to repay their creditors in 2002, while many advocated against foreclosing on distressed homebuyers in 2008.

Short-selling housing is hard and that contributes to the difficulty of arbitraging housing markets. It was hard for smart money to bet against booms, like Las Vegas in 2005. Short sales are made particularly difficult because of a lack of asset interchangeability (Nathanson and Zwick, 2013). Normally, a short is achieved by borrowing an asset from

someone else, selling that asset and then promising to buy it back. Enormous variation in characteristics across houses makes such a home short sale process almost impossible. Short-selling collateralized debt obligations or REITs are considerably easier, but these assets are quite different than ordinary homes.

Glaeser and Gyourko (2009) pointed out that other, even simpler, modes of arbitrage, such as delaying eventual purchases, are difficult to exercise because delaying a sure purchase introduces large amounts of risk in the portfolio of any average household. The limits of arbitrage in asset markets typically increase the possibility that prices will deviate from fundamentals (Shleifer and Vishny, 1997). The extreme limits of arbitrage in housing may make those deviations even larger.

A final peculiarity of housing markets is that the endogenous supply of new housing is so obvious that it cannot be ignored, while asset pricing economists routinely treat the supply of assets as being fixed. America built over 9 million new housing units between 2002 and 2006. In 2005 alone, Las Vegas permitted almost 40,000 new housing units and Phoenix permitted over 60,000. Hall (2003) emphasized that there is also a supply of Internet start-ups that surely should have influenced the willingness to pay for shares of existing companies in 2000, but many models of speculation treat the supply of assets as fixed.

During historic housing booms, price growth has typically been tempered in areas with long-run elastic supply. Between 1996 and 2006, nominal price growth was $3 \times$ higher in areas where housing supply is inelastic than in areas where housing supply is elastic (Glaeser et al., 2008). Yet, many elastic cities experience large price increases as well during housing booms. For instance, Las Vegas and Phoenix and cities in Florida and inland California—places with very elastic housing supply—witnessed many of the largest price increases in the nation between 2000 and 2006 (Nathanson and Zwick, 2013).

In our own research, we have taken different but complementary approaches to explaining these elastic price booms.⁷ Nathanson and Zwick (2013) argue that this phenomenon occurs in areas where housing supply is elastic today but will become inelastic soon. An example would be a growing city approaching a long-run development barrier. Investors speculate in the land markets about what the city will look in the future where new construction is difficult. This speculation leads to a boom and bust in the housing market, while the undeveloped land facilitates construction during the boom.

Glaeser (2013) took a different approach and argued that ignoring the impact of elastic supply is a pervasive error made by real estate speculators in the United States throughout the century. In many of these episodes, it took time for housing supply to come on line. When speculators don't realize supply is adjusting gradually, price booms occur that are followed by busts.

⁷ Gao (2014) and Sockin and Xiong (2014) also studied this phenomenon.

These theories differ in their predictions for what precipitates the bust. According to [Nathanson and Zwick \(2013\)](#), the bust occurs when optimistic land speculators learn that their rosy beliefs were incorrect. In [Glaeser's \(2013\)](#) argument, the bust happens when supply is finally able to adjust. The recent elastic booms fit the former theory better. Construction was high throughout the boom in these cities and not just at the end. Furthermore, consumer demand started to weaken in 2006 ([Mian and Sufi, 2010](#)), suggesting that belief revisions about household demand coincided with the beginning of the bust. [Glaeser's \(2013\)](#) mechanism better captures some historic episodes, such as the New York skyscraper boom of the 1920s, where construction was difficult to supply quickly due to new technologies.

Our theories both depart from standard theories of rationality but in different ways. [Glaeser \(2013\)](#) assumed that all market participants make the same common error of underforecasting future supply. This type of mistake has been suggested in other markets as well ([Hoberg and Phillips, 2010](#); [Greenwood and Hanson, 2013](#)). In many US cities, ignoring new local supply is largely irrelevant—the number of new homes is sufficiently small, and a general decision to ignore the supply side is reasonable. More generally, given that economists believe that teaching economics has some value added, perhaps we shouldn't be surprised that most Americans are not innately gifted at grasping the workings of supply and demand.

In contrast, [Nathanson and Zwick \(2013\)](#) assume only that a few well-capitalized investors are very optimistic. This aspect is a feature because it involves a smaller aggregate deviation from rationality. But it also poses problems, because homeowners must disagree with the optimistic valuations of houses and buy them anyway. It is certainly possible that homebuyers desperately wanted to move to Las Vegas in the 2000s at any price and bought housing expecting a capital loss. [Nathanson and Zwick \(2013\)](#) show that short selling of homebuilders that held large land portfolios increased during the boom, consistent with the existence of pessimists. But pessimistic homebuyers aren't representative of the common way of looking at the housing market.

11.2.4 Directions for future research

The user cost model itself is so well developed that further research on it is unlikely to yield rich results. Yet, there could be some benefits to developing a more comprehensive set of theorems about the connection between the underlying benefit stochastic process and the implied price series. For example, it would be nice to know what the available moments of price series imply for the possible range of unobserved local benefit ("rent") series.

By contrast, many of the unique aspects of housing haven't been integrated fully into financial models of housing markets, which often treat housing as just another asset. Despite considerable progress in these areas, we lack general theorems about how

housing price series are influenced by the fact that housing is bought and sold only by paying extremely high transaction costs and typically after considerable search. We still know too little about the connection between endogenous supply and price fluctuations.

One natural means of identifying the impact of housing-specific phenomena is to examine related markets that differ in small ways from housing. For example, commercial real estate shares many of the same attributes as residential real estate, yet it is typically held by dispersed owners but by asset managers with large portfolios. REITs can be sold short, while privately held developments cannot, so examining REITs allows us to assess the impact of yet another particular feature of housing markets.

11.3. EMPIRICAL REGULARITIES OF HOUSING DYNAMICS

At this point, we turn to the empirical regularities of real estate dynamics. In some cases, these figures are intrinsically interesting, but in many cases, we are interested primarily in whether these facts are compatible with the simple LAPM or user cost model. We are interested primarily in facts that seem somewhat anomalous and hence particularly worthy of future research. We chose to present facts after first discussing the benchmark LAPM or user cost model precisely because facts can only be anomalous unless they conflict with the predictions of a basic model.

We divide this discussion into four subcategories. First, we discuss excess variance in price movements relative to fundamentals. Since [Shiller \(1981\)](#), excess variance has been the biggest puzzle in asset markets more generally, and it is also a significant puzzle in housing. We then discuss short-term momentum and longer-term mean reversion of housing. We then turn to facts about quantities, rather than prices, by which we mean sales, vacancies, and the amount of new construction. Finally, we end with a brief review of the longer history of real estate movements. This last discussion draws heavily on [Glaeser \(2013\)](#).

11.3.1 Excess variance

Real estate is subject to fluctuations that are larger than they seem to be justified by fundamentals—just like other asset markets (e.g., [Shiller, 1981](#)). If these fluctuations are not caused by underlying fundamentals, then they fit the standard ([Fama, 1965](#)) notion of bubbles: price movements not caused by fundamentals. But this excess volatility is not uniform across time periods. Housing can be quite steady for many years, but there are periods when housing prices move far more dramatically in ways that are quite hard to square with fundamentals.

How big is the standard deviation of price changes for metropolitan area housing markets? [Glaeser et al. \(2014\)](#) report that the standard deviation of 1-year price changes in the United States ranges from \$2000 in the Sun Belt to \$13,300 in coastal cities over the 1990–2004 period. Five-year price volatility ranges from \$5400 in the Sun Belt to \$48,000 in coastal cities. Using the entire sample of cities for which FHFA data are

available, we estimate a standard deviation of price changes over the 1980–2004 period of slightly under \$10,000.

Are these numbers large or small? If they are benchmarked against changing rent values, they seem slightly large during the early time period. The standard deviation of rent changes using the REIT data suggests a standard deviation of about \$623 of annual rent changes. Thus, over the 1980–2004 period, the standard deviation of price changes is about $16 \times$ higher than the standard deviation of rent changes. But this difference is not wildly out of line with the price–rent ratio over the same period, so the price variance doesn’t seem all that excessive.

The same picture emerges when we compare price changes with the changes predicted by changes in income. Glaeser et al. (2014) compare price changes with those predicted by volatility in income, at least as measured by the average income of buyers found using the HMDA data. The volatility of prices over the 1990–2004 period is roughly similar to that predicted by income volatility. If BEA data on personal income across the entire metropolitan area are used instead of the income of new buyers alone, then the price volatility does appear to be significantly too high in coastal metropolitan areas. Head et al. (2014) similarly find variances that are in line with those predicted by a simple housing model.⁸

The real excess variance appears during periods, like that between 2001 and 2010, when prices temporarily explode (Wheaton and Nechayev, 2008). During this period, rents were relatively constant, and incomes didn’t move much. Yet, the standard deviation of price changes is over \$20,000. This is an extraordinary amount of variation across metropolitan areas that doesn’t appear to be related to any obvious changes in fundamentals. Figure 11.3 shows the path of rents and prices in Boston and San Francisco, where it is obvious the rents had peaked and were falling before the housing price boom crested.

As such, the data do not suggest that housing prices display the same ubiquitous excess variance found in many other asset classes. Instead, housing prices experience brief moments of extreme variance that punctuate longer periods of general stability. For example, the 1991–1996 period was an era of extraordinary price stability across America’s metropolitan areas. If this view is correct, then the puzzle is not to explain constant price variation, but rather periods when prices briefly explode and then tumble.

11.3.2 The shape of the cycle: short run momentum and long run mean reversion

For most observers, bubbles are embedded in real estate cycles and those cycles have features that go beyond excess asset price variance. A price series, $P_t = \hat{P}_t + \mu_t$, where

⁸ Other approaches find somewhat more excess volatility. Gelain and Lansing (2014) found excess volatility in price–rent ratios relative to “a simple Lucas-type asset pricing model.” Ambrose et al. (2013) found excess price volatility over a 355-year period in the Netherlands relative to fundamentals.

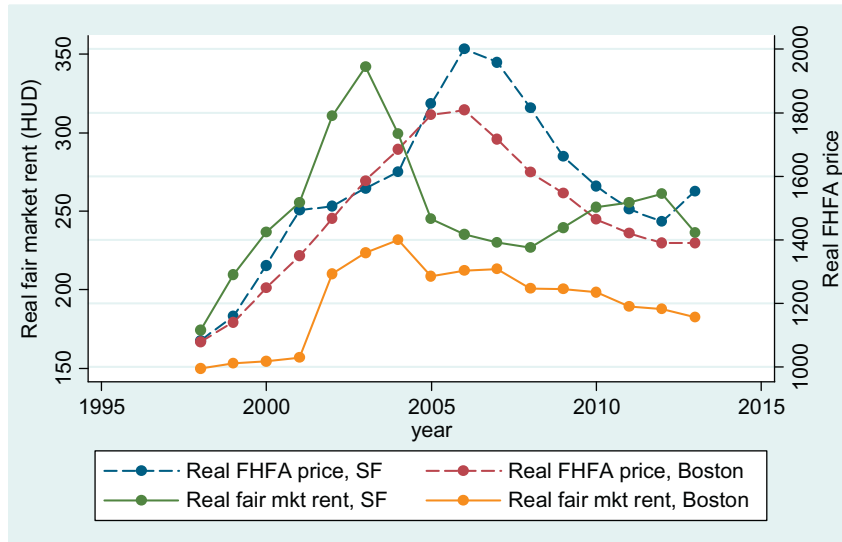


Figure 11.3 Price and rent growth for Boston and San Francisco. *Source: FHFA and HUD. Rents and prices in 2013 dollars.*

P_t is the realized price, $\hat{P}_t$ represents the price based on fundamentals, and μ_t reflects idiosyncratic white noise, independently drawn each day or every week, would fit Fama's (1965) definition of a cycle, but it would not capture the price movements that are often associated with major bubbles. Such a price series would be manic indeed, but not sustained, and manic high-frequency movements would have little impact on investment and probably not much on financial systems either.

In the popular view, influenced by classic descriptions such as Kindleberger (1978), the courses of bubbles have a defined shape. They begin with an early uptick, perhaps representing real good news or a little bit of early froth. That surge escalates, and during this growth period of the bubble, price growth escalates and the path of prices is convex. Ultimately, the bubble ends either in a violent crash or in a slow deflation.

Just like other markets, housing price changes display positive serial correlation at higher 1-year frequencies and mean reversion at lower frequencies like 5-year periods (Cutler et al., 1991; Glaeser et al., 2014). In normal periods, this momentum and mean reversion just seem like the normal course of affairs. During a great housing price event, such as the period between 2000 and 2012 for the United States or the 1980s boom and subsequent bust in Japan, the momentum and mean reversion define the shape of a bubble's path.

Over the 1980–2004 period, the one-period serial correlation of prices ranges from 0.75 in coastal metropolises to 0.6 in the Sun Belt, when we allow city-specific trends. The raw coefficient when prices are regressed on lagged price changes across the entire

sample from 1980 to 2012 is about 0.63 with or without city-specific trends. [Head et al. \(2014\)](#) report an even larger coefficient of 0.75.

These figures are remarkably high relative to the serial correlation of either rent or income changes. Both of these numbers are approximately 0.25. In the benchmark dynamic urban model of [Glaeser et al. \(2014\)](#), that serial correlation in fundamentals implies a price correlation that is essentially zero. The search model of [Head et al. \(2014\)](#) generates significant positive serial correlation in price but still far below the serial correlation seen in the data.

The positive serial correlation in 1-year price changes is particularly high during booms. [Table 11.1](#) shows the price correlation when price growth is regressed on lagged price growth year by year. As [Figure 11.4](#) also shows, the coefficient is over one at the height of the boom. This tendency of price growth to spiral is one of the most salient aspects of booms and one of the most difficult facts to reconcile with the simple models of housing price formation.

While housing prices show momentum at high frequencies, they mean revert at lower frequencies. Over 5-year periods, the correlation of price changes on lagged price changes ranges from -0.24 in the Sun Belt to -0.57 in coastal metropolises. [Figures 11.1 and 11.2](#) show the extremely strong 5-year mean reversion over the last decade.

This mean reversion is reasonably compatible with rational models, even if the magnitude of changes during this particular period is not. The dynamic urban model of [Glaeser et al. \(2014\)](#) essentially predicts mean-reverting prices similar to those seen in the data. This price mean reversion is predicted both by mean-reverting income processes and by new construction.

11.3.3 Quantities versus prices

There are also significant facts about the quantities of housing that dynamic models need to explain. Quantities both have a purely physical component—the number of houses being produced—and a market-related component—the number of houses on the market at a given point in time. Housing supply-related experts tend to focus on the former numbers. Experts on search and market dynamics focus on the latter number.

The volatility of construction is significant, especially in the Sun Belt. The standard deviation of annual permits in Sun Belt metropolitan areas is over 5000 units. In coastal metropolitan areas, the standard deviation falls to under 2000 units. While these numbers are significant, they do not appear to be particularly high relative to the number predicted by reasonable estimates of housing supply functions and the volatility of local income changes.

Permitting behavior shows remarkable persistence at 1-year periods as well. The serial correlation of permits over 1-year periods is significantly over 0.5. Again, this is compatible with the predictions of a simple rational model. If the costs of building increase with the amount of buildings, then it is sensible for booms to persist over multiple years.

Table 11.1 Correlation price growth on lag of price growth

(1)	(2)	(3)	(4)	(5)	(6)
Year	Log growth from a year ago	Coefficient on lag growth	SE on lag growth	r^2	Observations
1980	−0.03	0.00	0.15	0.00	79
1981	−0.04	0.29	0.10	0.07	116
1982	−0.01	0.20	0.12	0.02	131
1983	−0.01	0.03	0.06	0.00	135
1984	−0.01	0.08	0.05	0.02	146
1985	0.00	0.37	0.05	0.24	168
1986	0.03	0.59	0.05	0.42	187
1987	0.02	0.81	0.05	0.55	214
1988	−0.01	0.78	0.06	0.42	238
1989	0.00	0.33	0.04	0.16	297
1990	0.00	0.34	0.06	0.10	236
1991	−0.01	0.11	0.04	0.02	337
1992	0.01	0.21	0.02	0.19	349
1993	0.00	0.68	0.04	0.41	359
1994	0.02	0.89	0.04	0.56	366
1995	0.00	1.14	0.05	0.54	381
1996	0.02	0.17	0.02	0.15	392
1997	0.00	0.52	0.05	0.20	396
1998	0.03	0.17	0.03	0.07	397
1999	0.01	0.46	0.04	0.21	397
2000	0.01	1.19	0.08	0.33	399
2001	0.04	0.57	0.03	0.41	399
2002	0.03	0.44	0.03	0.35	399
2003	0.03	1.10	0.04	0.70	400
2004	0.04	1.11	0.03	0.73	401
2005	0.06	1.48	0.04	0.74	401
2006	0.07	0.83	0.04	0.57	401
2007	0.01	0.02	0.02	0.00	401
2008	−0.05	0.52	0.07	0.13	401
2009	−0.04	1.22	0.04	0.69	401
2010	−0.08	0.33	0.02	0.43	401
2011	−0.06	0.54	0.02	0.54	401
2012	−0.03	0.58	0.04	0.36	401

Source: FHFA

For 5-year permitting periods, there is substantial mean reversion, which appears less compatible with a rational model. In a sense, this seems to reflect a pattern where periods of overbuilding are followed by periods of underbuilding. Perhaps, this represents an example of excessive exuberance in construction.

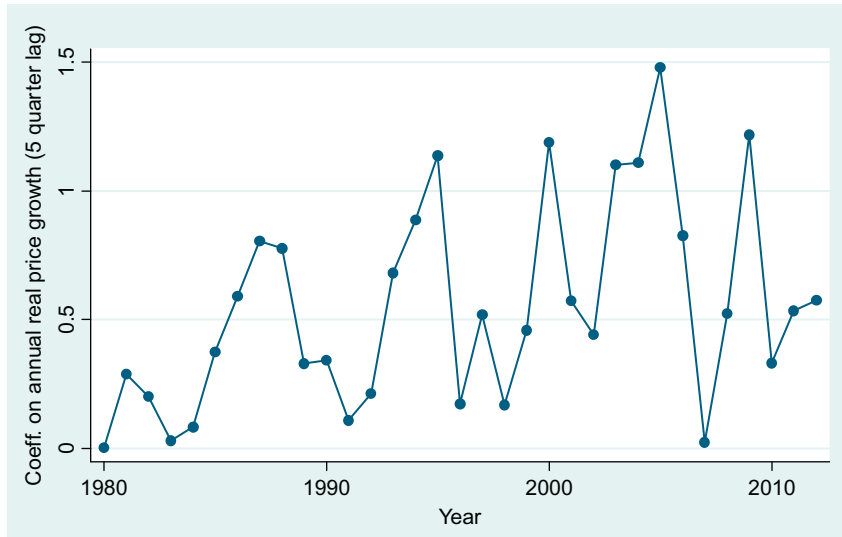


Figure 11.4 Real annual price growth on lag of real annual price growth. *Note: Dependent variable is first quarter annual growth. Source: FHFA.*

Head et al. (2014) also document that housing sales show significant persistence in the data. This is compatible with their model at 1-year frequencies, but there is too much sales persistence in the data, relative to their model, at longer frequencies.

Another particularly important stylized fact is the well-known relationship between sales volatility and price over the cycle. Markets are extremely active during booms and then dry up during busts. Some authors have interpreted this as evidence of nominal loss aversion on the part of sellers (Genesove and Mayer, 2001).

11.3.4 Bubbles in history

Given the rich history of real estate bubbles, it is useful to distinguish between two different classes of events: real estate fluctuations that are driven partially, at least initially, by new information coming from outside the real estate market itself and real estate fluctuations that appear to be driven primarily, sometimes almost exclusively, by events without the local housing market. Before 1980, real estate movements seem to have been typically associated with major uncertainty about external events.

The Chicago land boom of the 1830s, used by Shiller (2005) and others as an example of speculative mania, had a clear origin: the announcement that the state would fund a canal that would link the Great Lakes system with the Mississippi River system via the Chicago River. High cotton prices drove the Alabama boom of 1819 and high wheat prices helped justify Iowa land prices in 1910. In both cases, land buyers would eventually be burned, but price movements have justifications beyond merely the extrapolation of

price movements within the housing market itself. We aren't suggesting that these buyers were rational or that prices were themselves justified by fundamentals, but simply that these movements reflected more than buying based on past price movements.

By contrast, it is harder to see what external event could have motivated many of the price booms during the 2002–2006 period. The leading candidate is the proliferation of subprime technologies that expanded credit to low-income borrowers. This demand shock is not as clear as historic ones that involved the founding of cities. But [Mian and Sufi \(2009\)](#) presented evidence that less credit-worthy neighborhoods saw greater price increases during the boom. Investors may have had uncertainty about the long-run effects on housing demand of this credit expansion ([Nathanson and Zwick, 2013](#)).

Bubbles driven by exuberance about some external event can presumably take on any shape whatsoever, so long as prices eventually fall. A new announcement may spur an immediate price jump, which persists temporarily and then collapses, or the initial announcement may lead to ongoing price increases. An internally driven bubble must, however, display the positive serial correlation in price growth—the momentum that has come to almost define housing bubbles. The price growth itself is the news and the event that generates even more price growth. At the extreme, this creates the Alpine convex price pattern seen in places like Phoenix during the boom.

In this view, all bubbles are defined by large variance of price changes, relative to fundamentals and eventual mean reversion. Internally driven bubbles must also display positive serial correlation of price changes at higher frequencies. High levels of variance, positive serial correlation of price changes at high frequencies, and mean reversion at low frequencies are all well-known features of housing markets ([Cutler et al., 1991](#); [Glaeser et al., 2014](#)). It is more debatable whether these features reflect the workings of bubbles or changes in fundamentals such as interest rates.

11.3.5 Directions for future research

These empirical regularities need further documentation and they suggest a host of other topics for study. [Glaeser et al. \(2014\)](#) is only one paper on excess variance in housing prices. More research is badly needed, comparing excess variance in housing with other asset classes. If this empirical regularity is confirmed, then bubble models have an added challenge: explaining why excess variance is less common in housing markets than elsewhere. One might have thought that the prevalence of amateur homebuyers would create more excess variance relative to professionally dominated asset markets. Perhaps, the limits on homebuyers' purchasing resources act as a brake on irrational exuberance. [Nathanson and Zwick \(2013\)](#) emphasized that the marginal buyer in housing is likely to come from the middle of a large distribution of beliefs and hence the most optimistic beliefs will not drive prices. In asset markets, one or two large optimists can dominate markets, which may explain the greater level of excess variance.

The shape of the cycle also needs more research. The US housing facts may have been exhaustively documented, but we could badly use more equivalent information about housing price series elsewhere in the world and on commercial prices. Glaeser et al. (2014) suggest that new supply plays only a modest role in explaining the mean reversion of prices, but that is an artifact of particularly restrictive assumptions. Further research on the role that new supply plays in bringing markets back to earth would be valuable.

The correlations across housing markets remain a good topic for future research. These connections not only help teach us about how bubbles work but also teach us about the appropriate homeownership policy for individuals. Housing is more of a hedge when markets are more closely correlated, since movers who have experienced housing price appreciation are more likely to experience higher housing costs in any new market as well.

The economic history of housing bubbles is also a promising area for future research. There is an abundance of qualitative efforts documenting the course of the events and pointing out the apparent folly of the buyers. More quantitative research is needed, especially if it can document the connection between the boom and the fundamentals of that era.

11.4. RATIONALIZING THE SEEMINGLY IRRATIONAL: SEARCH, HETEROGENEITY AND AGENCY PROBLEMS IN CREDIT MARKETS

We now turn to the economic approaches that have been used to help understand housing price movements that do not seem to be in line with the simple version of the LAPM discussed above. In this section, we discuss explanations that assume rational buyers. In the next section, we focus on less rational theories. We split these rational theories into three groups: “search, learning, and momentum,” “changing credit conditions,” and “rational bubbles and agency problems.”

11.4.1 Search, learning, and momentum

We have already emphasized that housing markets are, in reality, highly decentralized markets with a great deal of heterogeneity. Search is a major feature of this market and there is no way that buyers or sellers can simply observe the current price of housing. Decentralization therefore is related to the level of information in the housing market, and those authors who suggest that features like momentum can represent learning are either implicitly or explicitly relying on the decentralization of the market. Several authors such as Wheaton (1990) and Krainer (2001) had shown that search and learning can help explain key features of housing price dynamics. Han and Strange (Chapter 13) surveyed the literature on search models in housing.

We will turn to limited information later, but even with perfect information, decentralized markets, especially when combined with some price stickiness, can lead to

staggered price responses to shocks which can create momentum in realized average prices. In [Head et al. \(2014\)](#), individuals enter into the housing market and choose whether to search for a home or just to rent. Some homeowners decide whether to sell or rent their houses. An increase in local area income causes an influx into the city. This influx causes a conversion of vacant homes to rental units, because matching renters is far easier than matching owners, and it also causes a gradual buildup in the number of would-be buyers, not all of whom are immediately matched with houses to buy.

This buildup of buyers causes market tightness to increase over time, which is also a feature of [Genesove and Han \(2012\)](#).⁹ The rising number of buyers to sellers causes prices to increase even further. Eventually, new construction brings prices back down. This paper does an admirable job of fitting a number of features of the housing market, including patterns of sales and vacancies. But most notably, it manages to predict price momentum due to increasingly tight markets because of unmatched buyers.

Yet, this model does not really predict bubbles. It predicts that a positive shock might take several periods to work its way through the system, but not that the price movement might be many times greater than that implied by fundamentals. As such, this type of search model seems like it can clearly be part of the explanation for the structure of housing bubbles, especially their price momentum, but not for the bubbles themselves.

[Guren \(2014\)](#) provides an alternative search model that can also generate sticky adjustments to shocks. [Guren \(2014\)](#) assumes that sellers change their prices only every other period—there is an ad hoc stickiness in the model. As a result, only one-half of sellers change their prices during each period and the other half keeps their prices fixed. Sellers who price high, relative to the market, don't attract many searchers to their homes. This means that sellers are slow to react to upward demand shocks, because quick leaps to the equilibrium prices will make it hard to sell the property. The slow equilibration process can lead to momentum, even if there is perfect information about the state of demand.

There are limitations to the power of this type of argument. First, it is a better argument for upside momentum than downside momentum. Pricing high is particularly costly because it deters buyers from even visiting the house, but pricing low has no equivalent cost, which means that there is no strong force preventing sellers from quickly adjusting the prices downward.¹⁰ Second, there is nothing in this argument that can lead to excess volatility, without some outside force such as exogenous beliefs. However, once there is an external forcing process creating excess volatility, as in [Piazzesi and Schneider \(2009\)](#), then a decentralized market can exacerbate the impact of that outside force.

⁹ In that paper, tightness overshoots, which amplifies the fundamental shocks.

¹⁰ [Novy-Marx \(2009\)](#) provided a more symmetric amplification mechanism that operates through bargaining and the flow of entrants. This stimulating paper treats the flow of buyers as the primitive, rather than rents or the value of living in the locale. As such, it is hard to determine how these results relate to the excess variance that we are discussing here.

In these models, search primarily refers to matching the idiosyncratic features of the house with the tastes of idiosyncratic buyers, not learning about market fundamentals. Yet, it is surely true that buyers and sellers also gradually learn about market fundamentals through the search process. Even if they observe listed prices before the search process begins, search enables them to assess unit quality more thoroughly. Time spent on the market also acquaints buyers with actual sales prices for units that they have observed. This represents a form of learning about the state of the market.

There is ample evidence of ignorance and learning within housing markets. [Clapp et al. \(1995\)](#) provided evidence on the existence of various forms of rational learning. [Levitt and Syverson \(2008\)](#) illustrated that home sellers appear to be somewhat uninformed about the state of the market. Yet, while incorrect beliefs certainly have the power to move markets in many ways, it is unclear if rational learning can really generate particularly large movements in housing prices.

Rational learning suggests that buyers have formed some Bayesian estimate of the state of fundamentals and that estimate changes over time as new evidence trickles into the market. Generally, rational ignorance leads to less variance, not more, because individuals recognize how little they know. [Shiller's \(1981\)](#) variance bound test essentially lives off this point. Moreover, standard learning models still won't have predictable errors, so regular momentum or mean reversion in beliefs is incompatible with standard Bayesian learning.

To make this point precise, we return to the standard user cost argument in which P_t equals $R_t + \frac{E(P_{t+1})}{1+r}$. We assume that $r=0$ and $R_t=0$, so in this case, P_t equals $E_t(P_{t+1})$. It is impossible for anything known at time t to regularly predict the updating between t and $t+1$, so $\text{Cov}(P_{t+1} - E_t(P_{t+1}), P_t - P_{t-1})$ must equal zero. This fact implies that highly rational learning on its own will not help us to understand momentum or mean reversion. Learning by individuals who are rational but ignorant seems unlikely to generate—on its own—any of the three salient features of price dynamics that we have discussed above.

This claim does not mean to suggest that semirational learning, employing rules of thumb (as in [Shiller, 1999](#)) or extrapolating, will not help explain housing price patterns. This seems quite likely, but we will address these issues later as we turn to less rational models.

11.4.2 Changes in credit conditions

Perhaps, the most popular “rational” model of housing price changes is that they reflect changing credit market conditions, which can mean low interest rates or permissive lending or both. Typically, some external factor, like a glut of Chinese savings, is justified to explain the credit market changes, which allegedly explain housing booms and busts. Yet, the common view that housing crisis was caused by subprime mortgages and a global lending glut has difficulty making sense of the heterogeneity across the US credit markets are

national or global, so if easy credit was the cause, then why did Houston experience almost no boom and bust while the cycle in Las Vegas was particularly dramatic? The heterogeneity across countries is also hard to explain if the boom is solely the result of a massive flood of global lending. These differences are also incompatible with a common, credit-caused demand shock interacting with heterogeneous supply (Davidoff, 2013).

Certainly, there is considerable evidence that easier credit did induce buying among subprime households in the years before 2007 (e.g., Mian and Sufi, 2009), and lower interest rates are generally associated with higher housing prices (e.g., Poterba, 1984). The rise of subprime lending certainly the boom establishes little about causality because the lending itself may be a reflection of overoptimistic beliefs, which are causing both phenomena.¹¹ Mian and Sufi's (2009) contribution is a tight focus on causal inference, which has considerably increased our confidence that subprime lending boosted prices, yet even in their work, subprime lending seems to only explain a modest fraction of the rise of housing prices.

Moreover, it is far from clear whether volatile interest rates can create massive housing price fluctuations in a purely rational model or at least a model without bubbles. We explore rational bubbles with overly cheap credit in the next section. Here, we briefly discuss the impact of credit market changes in the standard LAPM.

One standard version of the LAPM implies that $P_t = \frac{(1+r)R_t}{r-g_R}$ or price equals the $1+r$ times the flow value of housing divided by the difference between the interest rate and the growth rate of fundamentals. This formula would seem to imply an extremely tight relationship between prices and interest rates, especially in a high growth environment. This formula was used during the boom to justify extremely high prices.

But while this formula is correct in a static interest rate environment, it is not correct in a world in which interest rates are dynamic. For if interest rates mean revert, then buyers during periods when interest rates are low should expect to sell when interest rates are higher and vice versa. Rational expectations about the changing value of interest rates should cause buyers to be less responsive to changes in interest rates than a naive application of this formula would suggest.

Unfortunately, dynamic interest rate models do not yield easy closed-form solutions for housing prices, but Glaeser et al. (2013) simulated rational prices in a world in which interest rates followed a mean-reverting Cox–Ingersoll–Ross diffusion process. They found that the implied semielasticity of prices with respect to interest rates (the derivative of log prices with respect to a change in the interest rate) can drop by as much as two-thirds when the mean reversion of interest rates is taken into account. So, while the naive model predicts a relationship between interest rates and prices that is large enough to

¹¹ A similar argument can be made about the correlation between investors buying homes and the boom (e.g., Haughwout et al., 2011). The investors seem more likely to be a reflection of market enthusiasm than to be an independent cause of rising prices.

explain much of the price increase between 2001 and 2006, a forward-looking model cannot. Moreover, the relationship between interest rates and prices implied by the forward-looking model does a better job of fitting the historical relationship between prices and rates over the long run.

There are at least three other theoretical forces that tend to reduce the connection between interest rates and housing prices. Elastic supply will tend to reduce any connection between demand-side forces, like credit conditions, and prices. The ability to refinance in the future will similarly make current conditions less important. Finally, if buyers discount the future using their own internal discount factor, rather than the market rate of interest, then that market rate of interest is predicted to have a smaller impact on price.

A final issue is the exact timing of interest rate changes over the period from 2003 to 2010. Prices continued to rise despite the slight rise in interest rates after 2004. Prices fell dramatically despite declining interest rates after 2007. The observed price changes can still be explained as reflecting changing credit market conditions, since interest rates do not reflect the probability of actually getting a loan. As discussed above, the global credit glut view also has difficulties explaining the heterogeneity within the United States and across the world.¹²

We do not mean to suggest that credit conditions are irrelevant in housing markets or in explaining bubbles, but simply that the relationship is not the simple one suggested by the LAPM. The LAPM is based on assumptions of rationality yet applying such a static model to a dynamic interest rate process assumes buyers are myopic, not rational. To provide an intellectually coherent framework, we must more fully embrace bounded rationality, and we will turn to that shortly.

Interest rates are only one aspect of credit markets. Mortgage approvals and down payment requirements may also shape housing price increases, and foreclosures may exacerbate housing price decreases. The simplest model for understanding why such credit conditions matter is to assume a variant of the LAPM with individual heterogeneity to allow for a downward sloping demand curve for housing. In such a model, the value of R_t differs by consumer and is denoted as $R_t(i)$, and consumers continue and are willing to spend as much as $\frac{(1+r)R_t(i)}{r-g_R}$. The consumers with the highest valuations will be inframarginal consumers, and we assume that there is a distribution $G(R_t)$ of these valuations across the set of feasible buyers. If N_S denotes the number of homes being sold and N_D denotes the number of potential buyers of whom only a fraction θ can manage financing, then N_S , the supply of homes, equals $(1 - G(R_t^*))\theta N_D$, the demand for homes, where R_t^* reflects the valuation of the marginal buyer.

¹² Mian and Sufi (2009) documented that the price boom was bigger in areas where credit seems to have been more constrained ex ante, which suggests that there is a geographic component of the easy credit hypothesis. Yet, that does not imply that this hypothesis can explain the large heterogeneity across metropolitan areas.

In this model, an increase in the availability of credit, captured by the parameter θ , will act to increase aggregate demand. The derivative of price with respect to θ is $\frac{(1-r)(1-G(R_t^*))}{(r-g_R)\theta g(R_t^*)}$, and the elasticity of price with respect to θ equals $\frac{1-G(R_t^*)}{R_t^* g(R_t^*)}$, which is exactly the same as the elasticity of price with respect to the number of potential buyers. In principle, this type of calculation can justify a robust relationship between approval rates and prices, just as the LAPM seems to justify a robust relationship between interest rates and prices, but some of the same issues arise. We are assuming a permanent change in approval rates, but periods of easy credit tend to be followed by periods of tight credit.

If approval rates rise temporarily, then buyers should expect future approval rates to fall. Buyers should anticipate this future drop, and this should cause buyers who expect to resell to become less sensitive to current credit availability. Mean reversion should likewise occur for down payment requirements and this will also dampen the connection between prices and credit market conditions. A fully specified dynamic model linking time-varying approval rates to housing prices has not yet been written, but it would be a worthy addition to the literature.

We have implicitly assumed that mortgage approval is unrelated to the individual's valuation of the house and that those who are denied credit are a random subsample of the buying population. That need not be the case. Individuals with financing may be richer and willing to pay more, or, conversely, individuals who were initially denied financing might be particularly eager to buy housing, perhaps because they are more risk-taking. One way to understand the striking findings of [Mian and Sufi \(2009\)](#) on the connection between subprime mortgage affordability and increased prices is that more lenient lending made it possible for high-risk, but highly optimistic, buyers to enter the market.

[Stein \(1995\)](#) provided an alternative mechanism through which credit markets influence price fluctuations. If individuals are largely credit-constrained, then a price movement upward creates significant capital gains for existing owners, which then enables them to purchase even larger houses. By contrast, a price drop will mean that these buyers are essentially locked into their existing homes and cannot buy elsewhere ([Ferreira et al., 2010](#)). There is a housing price multiplier because past housing price appreciation is providing the cash to fuel future housing price purchases.

While this section has focused primarily on the credit market causes of the price boom, there is also a significant literature connecting credit markets with the price bust and the adverse consequences of that bust. Perhaps, the most common idea is that housing busts generate defaults, and defaults then have adverse consequences on the health of the housing market. [Palmer \(2013\)](#) documented compellingly that the wave of defaults that followed the 2007 bust were the result of falling prices, not the composition of buyers who received credit immediately before the bust.

[Campbell et al. \(2011\)](#) documented both that forced sales, such as default, receive lower prices but that they also seem to lower the prices of nearby homes. [Fisher et al.](#)

(2013) found that the negative effect of condominium foreclosures on their neighbors seems localized to the particular address. Mian et al. (forthcoming) found that housing prices dropped more after 2007 in states that made it easier for lenders to foreclose on housing.

Why would defaults cause housing prices to drop? Guren and McQuade (2014) highlighted three effects: Distressed sellers tend to be more impatient and sell for less, buyers will become more choosy because they have more of a chance of interacting with a distressed seller, and homeowners who default are themselves priced out of the market. Taken altogether, they estimate that foreclosures may have exacerbated the price decline by as much as 50%.

Credit institutions may also impact the consequences of housing market downturns. Somewhat obviously, the impact that housing busts have on financial institutions is due primarily to the fact that these institutions are providing credit for real estate investors, sometimes directly and sometimes by holding real estate-related securities, such as collateralized mortgage obligations. More subtly, housing prices appear to have a substantial wealth effect, and falling housing prices have led to substantial reductions in consumer spending and associated economic activity (Mian et al., 2013).

The compelling work of Mian and Sufi (2014) documents a wealth effect of housing that seems far stronger than the wealth effect of stock prices. Case et al. (2005) found a similar result. One explanation for this phenomenon is that stocks are owned by wealthier people who are not credit-constrained, while homes are owned by people who often are credit-constrained. Relaxing a credit constraint can easily have a far more powerful effect on spending than making an unconstrained individual wealthier, which may in turn explain why housing booms and busts appear to exert an outsized influence on the larger economy.

We now turn to rational bubble models in which credit market conditions can engender the possibility of dramatic price swings.

11.4.3 Agency, underpriced default options and rational bubbles

We have already discussed the two difficulties facing the models of rational bubbles in real estate: the standard violation of the transversality condition and the less standard problem of predicting an essentially unbounded supply of new housing. These difficulties become far less severe if homebuyers are charged too little for the possibility of defaulting on their mortgage.

We now more formally model the mortgage process so that buyers initially pay only a down payment fraction “ d ” of the purchase price of the house. At the end of the period, they either sell the house and repay the mortgage or default, and we assume that default carries a cost of “ z ” to capture the fact that many individuals fail to default even when that default would seem to be in their own interest:

$$R_t = dP_t - \beta E \text{Max}(P_{t+1} - (1+r)(1-d)P_t - z), \quad (11.2)$$

where r is the market discount rate, possibly adjusted to address default risk. It is natural to assume that $\beta \leq \frac{1}{1+r}$, since homebuyers presumably always have the ability to earn the market rate by savings (or by just not borrowing money) but they can't always borrow freely and hence they may value future dollars at a rate above the market rate. To clarify issues, we assume that there is no uncertainty in R_t , so in the absence of bubbles, prices will remain fixed, so $P = \frac{R}{d(1-\beta) + \beta(1-d)r}$, which is the flow value divided by a weighted average of effective discount rates. The weights depend on the extent to which the loans are self-financed.

In this case, a deterministic bubble would satisfy

$$B_{t+1} = \left(\frac{d}{\beta} + (1+r)(1-d) \right) B_t, \quad (11.3)$$

where B_t denotes the “bubble” component of the price. This equation creates the same challenges involving the transversality condition and infinite housing supply. While deterministic bubbles don't seem to be feasible, stochastic bubbles can exist if they lead to default in negative states of the world. We consider a bubble with a value of B_t that takes on a value at $t+1$ of B_{t+1} with probability $1-\nu$ and 0 otherwise. We simplify by assuming that $P_F = \frac{R}{d(1-\beta) + \beta(1-d)r}$ is the fundamental value in all cases so that the price with the bubble equals $P_F + B_t$. We assume further that if the bubble bursts, the homebuyer defaults. The technical condition for this to be optimal is that $P_F(d(1+r) - r) + z < B_t$. The equilibrium condition for a stochastic bubble is that

$$B_{t+1} = \frac{(d + (1-\nu)\beta(r - d(1+r)))P_F - R + \nu\beta z + (d + (1-\nu)\beta(1+r)(1-d))B_t}{(1-\nu)\beta}. \quad (11.4)$$

In the extreme case where $P_F = R = z = 0$, so the bubble alone remains, this simplifies to

$$B_{t+1} = \frac{d + (1-\nu)(1-d)(1+r)\beta}{(1-\nu)\beta} B_t \quad (11.4')$$

The bubble's value increases but only if it fails to burst, that is, with probability $1-\nu$, hence, the discounted value at time $t+j$ equals $B_t(d + (1-d)(1-\nu)(1+r)\beta)^{t+j}$. If $(1-\nu)(1+r)\beta = 1$, so that the interest rates are set so that the discounted expected value of a loan of one dollar is equal to one, then $\frac{B_t}{(1-\nu)\beta} = B_{t+1}$ or so $B_t = \beta^j E(B_{t+j})$ bubble maintains its expected value.

The standard transversality condition is that $\lim_{j \rightarrow \infty} \beta^j E(B_{t+j}) = 0$, and so this structure violates this condition. As the transversality condition holds for the fundamental part of the housing value, violating the transversality condition implies that people expect that in the long run, the bubble component alone will determine the home's value. [Diba and Grossman \(1988\)](#) argued that such explosive behavior implies that rational bubbles are

impossible, especially when assets can be supplied into the market. Santos and Woodford (1997) similarly suggested that the conditions that admit bubbles are “fragile.”¹³ Influenced by these papers, we also believe that rational bubbles seem implausible when they imply extraordinarily high future housing values.¹⁴

Bubbles become far more plausible if lenders underprice default risk, which might reflect agency problems with banks or irrationality by lenders. If interest rates are set too low, so that $1 > d + (1 - d)(1 - \nu)(1 + r)\beta$, then lenders aren’t fully compensated for the threat of default. In this case, the expected value of the bubble decreases over time and the transversality condition is not violated. Housing prices will rise with the bubble, but the rises may be less extreme. Still, there is a nonzero probability that housing prices will become extremely high and this should predict an enormous supply response.

There are conceivable ways of salvaging the model with endogenous supply. Perhaps, builders are sufficiently risk-averse, and it takes enough to build so that anticipating the possibility of a burst necessarily leads to limited production. This would be most plausible if ν is high so that the probability of default is quite high. Still, the fact remains that construction was extremely high in Phoenix, Las Vegas, and Miami during the recent boom, which makes the case for this type of rational bubble model more difficult to make.

Moreover, since a fully priced risk would require the bubble to rise quickly enough to violate the transversality condition, this finite price bubble requires the underpricing of risk. This underpricing does seem to be a regular feature of booms, but why does it occur? One explanation is that lenders are irrational, for some reason, but that violates the spirit of this section, which is to explain real estate bubbles with rational models.

The alternative explanation, which has been offered by many observers, is that there is an agency problem within the lending sector (Green, 2008; Diamond and Rajan, 2009), perhaps because of federal deposit insurance (Demsetz et al., 1997). In principle, this could occur within banks themselves even if those banks have no deposit insurance and hold mortgages on their own balance sheets. The CEO hires agents to make loans at an appropriate interest rate. Their compensation is increasing with the number of loans or the average interest rate paid, but limited liability precludes severe punishment in the state of the world when the bubble bursts. The result is that agents compete and the market rate of interest charges too little for default risk. Pavlov and Wachter (2006) argued that if some bankers are underpricing default risk, competition may push all bankers to underprice default risk.

¹³ These arguments are different than the alternative view that bubbles cannot exist because they will be arbitrated away. Abreu and Brunnermeier (2003) made a compelling case that this is unlikely in financial markets and the power of arbitrage is even weaker in housing.

¹⁴ Kivedal (2013) found evidence for the existence of a bubble in the housing market between 2000 and 2013 but concluded that it is more likely to be irrational than rational. Giglio et al. (2014) rejected infinitely lived bubbles in the United Kingdom and Singapore by comparing very long-run housing rental contracts to infinite ownership rights.

The larger debate over agency risk in lending concerns the creation of mortgage-backed securities. In this version of the hypothesis, mortgage originators have little incentive to screen for risk because they pass those risks downstream to the eventual security holder who has little ability to appropriately learn the truth. The presence of mortgage insurers, such as Fannie Mae and Freddie Mac, further decreases the incentive to price appropriately for risk. [Purnanandam \(2011\)](#), for example, found that banks that originated loans primarily to distribute them to downstream investors generated “excessively poor quality mortgages.” [Piskorski et al. \(2010\)](#) found a link between securitization and default, which [Adelino et al. \(2010\)](#) hotly disputed.¹⁵

Of course, this type of model still raises questions. If the risk comes from a systemic bubble, rather than idiosyncratic risk, then investors should be able to understand that they are subsidizing the homebuyers. The nature of the rational bubble is that its features are widely known and homebuyers supposedly understand the risks perfectly. If they understand the risks, then so should the ultimate investors. Agency problems typically occur when there is local information that can’t be observed by the principal, but there is nothing local about a system-wide bubble.

Perhaps, the best explanation for why systemic risk might be subsidized is that the government is bearing the tail risk. If the public sector is bearing the tail risk for political reasons, then this still could lead to underpriced default risk that encourages the emergence of rational bubbles. Nonetheless, this theory would still have to explain why private mortgage insurers led the way into the subprime lending that would be at most risk from a bubble bursting.

A slightly different alternative to this purely qualitative rational bubble model is to use the option pricing model of [Krainer et al. \(2009\)](#) and ask, as in [Glaeser \(2013\)](#), how much housing prices would increase if borrowers were given an non-priced default option. The model requires a switch to continuous time, so we must assume that the flow value of the house equals $r(t)$, where r follows a geometric Brownian motion. The drift is g_R ; the variance is σ^2 . Individuals discount at a rate but pay in interest on their mortgage each period or r times the outstanding debt or $r(1-d)P$, where P refers to the initial purchase price. The borrowers are only paying the interest on the mortgage. We assume costless default and the purchase price of the house will then satisfy

$$P_t = \frac{\rho}{\rho d + r(1-d)} \left(\frac{z^z (\rho - g_R)^z (r(1-d)P_t)^{1+z}}{\rho^{1+z}} R_t^{-z} + \frac{R_t}{\rho - g_R} \right), \quad (11.5)$$

¹⁵ [Piskorski et al. \(2010\)](#) found that securitized loans are more likely to be foreclosed and less likely to resume making payments, suggesting that renegotiation may be less likely with securitized loans, perhaps because of agency problems. [Adelino et al. \(2010\)](#) found little evidence of renegotiation in nonsecuritized or securitized loans and no difference in loan modification between the two samples.

where $z = \frac{g_R - 0.5\sigma^2 + \sqrt{(g_R - 0.5\sigma^2)^2 + 2\sigma^2\rho}}{\sigma^2}$. The first term in brackets represents the value of the default option; the second term reflects the value of the house without any possibility of defaulting. By calibrating the model, Glaeser (2013) estimated just how much an unpriced default option could contribute to higher housing prices during historic booms. This approach assumes that the market rate and the private discount rate are the same and then calculates the value of the first term in the expression. Typically, this represents less than 17% of the price for more recent price swings, which is surely an overestimate since it assumes that interest rates incorporated none of the cost of potential default.

This calculation again illustrates that if credit markets are responsible for the extreme volatility of housing bubbles, then it seems more likely they do so by making conditions ripe for bubbles. This view suggests that there is not an automatic link between housing prices and easy credit, but rather that easy credit is a necessary—but not sufficient—cause of extreme price volatility.

11.4.4 Directions for future research

All of the topics discussed above contain major open questions. Search in housing is currently a hot topic, and recent papers have yielded promising results. Yet, we are still unsure whether some form of essentially rational search model can deliver all of the stylized facts discussed in Section 11.3. It is clear that these models can deliver momentum, but excess variance does not spring so readily from such models. Moreover, it is yet to be seen whether these models can explain the cross-sectional variation in housing market behavior.

The role of credit markets in housing cycles is also far from settled. Indeed, the best interpretation of Glaeser et al. (2013) is not that easy credit didn't cause the cycle, but that the confident view that it did cause the cycle has little solid support. More empirical work is surely needed, especially taking advantage of international data where they might be usable independent variations in credit conditions. More theory is also needed, particularly analyzing the role of credit in near-rational models.

The connection between bubbles and underpriced default risk is sketched here, but a full treatment of that topic would also be useful. Is it really plausible that agency problems could, with no irrationality, generate underpriced default risk that is large enough to produce large rational bubbles? Can such a model deliver other predictions that could be tested?

11.5. A MENAGERIE OF MODEST MADNESS: BOUNDED RATIONALITY AND HOUSING MARKETS

We now turn to less than rational models that have also been used to investigate real estate fluctuation, and typically, those models involve buyers who hold excessively optimistic

beliefs about future housing price growth. The Case and Shiller surveys of recent buyers suggest that such beliefs certainly exist. For example, [Case and Shiller \(2003\)](#) reported that buyers in Orange County, California, in 1988 expected prices to rise by 14.3% per year over the next 10 years, and in 2003, they expected prices to rise by 13.1% annually over the next 10 years. [Case et al. \(2012\)](#) reported that expected price growth over the next 10 years from buyers in Middlesex County, Massachusetts, fell from 10.6% in 2004 to 3.1% in 2012.

These striking survey results need a grain of salt, for they are surely muddled by innumeracy and wishful thinking. Yet, even if the numbers appear excessive, they surely capture an important reality. Many buyers during booms seem to have expectations that are wildly optimistic and often at odds with the views of economists and the experience of longer-term price trends.

We begin this section with exogenous belief models, in which individuals for some reason have beliefs that are unduly optimistic. As models in which beliefs are entirely flexible and determined outside of the model have the potential to “explain” any housing event, they are not particularly compelling as theories without some added ingredient. Exogenous belief models are typically used to illustrate some other point about housing markets, such as the role of search or endogenous housing supply. The second class of models that we consider are extrapolative beliefs, and we consider both a naive version where extrapolation is merely assumed and a more sophisticated version in which extrapolation emerges out of cognitive limitations. We then turn to the broad class of models with cognitive limitations that involve shorter time horizons and simple models of housing price formation. We end with a discussion of social learning.

11.5.1 Exogenous and heterogeneous beliefs: search, endogenous housing supply, and land acquisition

The simplest way to get a bubble is just to assume that individuals are unduly optimistic. Using the standard pricing formula where P_t equals $R_t + \beta E(P_{t+1})$, we can justify almost any price process imaginable by assuming differential values for $E(P_{t+1})$. Similarly, if pricing follows a growth formula like $P_t = \frac{R_t}{1 - (1 + g_R)\beta}$, then exogenous changes in the belief about future rent growth will naturally cause fluctuation in the price and anything is possible. Since anything is possible, the exogenous optimism model on its own essentially fails as social science, for it cannot be rejected by the data, and a model that cannot be rejected cannot be tested. Essentially, in the move from perfect rationality to exogenous beliefs, we have moved from a theory with predictions that are too narrow to a theory that has predictions that are too broad. If we believe that beliefs are not perfectly rational, and we want to test that hypothesis, then we must assume a particular structure for beliefs that can be tested.

While the assumption of exogenous beliefs cannot be tested on its own, it does provide testable implications when nested in a larger model. Two examples of this structure

are Glaeser et al. (2008) and Nathanson and Zwick (2013). Glaeser et al. (2008) assumed that bubbles are generated by random increases in the buyers' valuation of homes. They examined how this exogenous shift in demand interacts with supply.

During a boom, holding the degree of irrational surge constant, places with more inelastic supply will have greater increases in price and lower increases in quantity. This claim merely makes the point that the logic of Economics 101 continues to operate even if the demand curve is being shifted by irrationality. This point follows Becker (1962). After the bubble, the impact of supply elasticity is ambiguous. If supply is extremely inelastic, then the bubble will have had no impact on quantity during the boom and hence little impact on prices after the boom is over. This may reflect the reality of northern California or Massachusetts. In extremely elastic places, bubbles cause explosions in homebuilding, but the elasticity of housing supply itself mutes the impact of overbuilding ex post, as long as the area continues to build at all.

Nathanson and Zwick (2013) also explore exogenous belief shifts, but their focus is on heterogeneity. They assumed that individual investors have different prior beliefs about the value of real estate in a different area. They ruled out the possibility that beliefs will converge through a learning process. Optimists buy up all the land and the rental housing. But owner-occupied housing stays dispersed among residents of all beliefs because some residents prefer owning over renting, and they have diminishing marginal utility of housing. Optimistic beliefs influence house prices most in areas with ample land or rental housing. Prices are less prone to bursts of extreme optimism in housing markets that only have owner-occupied housing and no undeveloped land.

This observation explains why many elastic areas, which are those with undeveloped land, have such large house price booms. Nathanson and Zwick (2013) show that land price increases capture nearly 100% of the dispersion in house price increases across metro areas. Furthermore, several US public homebuilders acted like speculators by taking large positions in the land market between 2001 and 2006 and then suffering large capital losses. Short selling of homebuilder stock rose dramatically during this period, providing evidence of pessimists who disagreed with the homebuilders' high valuations of land.

11.5.2 Extrapolative beliefs

Perhaps, the most popular alternative to rationality in housing markets (and perhaps finance as well) is extrapolation, occasionally called momentum trading or backward-looking investment. Clapp and Tirtiroglu (1994) is an early example of this assumption in housing economics. In the real estate context, Glaeser (2013) referred to extrapolators as Gordonians because of their blind use of the Gordon growth formula. The principle is simply that investors use a formula like $P_t = \frac{R_t}{1 - (1 + g_R)\beta}$, and they use the recent growth rate in past prices to infer the growth rate in fundamentals. There are two ways of generating this type of behavior. One option is to assume that these erroneous beliefs arrive

exogenously in a small share of the population. A second option is to derive these beliefs as the result of primitives.

The choice between these two options is somewhat connected to the decision about whether to assume that the extrapolative bias is ubiquitous or particular to a few odd eggs. Finance has a tradition going back to DeLong et al. (1990) of assuming that irrationality is limited to a small share of the market and then looking at what impact this irrationality would have on market-wide prices. The appeal of this approach is that irrationality can be assumed for only a few, preserving the possibility that most of us are rational.

Barberis et al. (2015) documented how a small number of extrapolative buyers can move prices in financial markets. Piazzesi and Schneider (2009) provided an elegant example of this tradition transplanted into housing markets, which makes the point that if there is only a modest share of the stock of renters with extrapolative beliefs, that small share of the stock can become a large share of the flow of new home purchases, since the flow of purchases is only a small share of the stock.

Guren (2014) follows a similar tack of examining the impact of a minority of backward-looking investors (i.e., extrapolators) in his model of search and slightly sticky prices. He finds that his model also exacerbates the power of a small amount of irrationality, allowing it to propagate throughout the system. In particular, small numbers of extrapolators create significant positive serial correlation in housing prices.

While Piazzesi and Schneider (2009) did persuasively show that a small number of optimistic buyers can make a boom, we think that the 2002–2006 boom is best seen as a far more widespread burst of optimism. We agree that boom-level prices may not have reflected the majority opinion and surely did not reflect the views of a sizable minority of housing market skeptics. Yet, millions of Americans thought that these high prices made enough sense to purchase houses. Perhaps, these homebuyers were extrapolating recent price increases, or perhaps, they were passive and not evaluating house prices carefully because of long horizons. The available survey data (Case et al., 2012) suggest that optimistic beliefs were ubiquitous among buyers during the boom period, although that comes with the usual caveats on survey data.

A moderate deviation from strict rationality can take the form of either large deviations from rationality for a small number of people or small deviations from rationality for much larger groups. While some economists may be comfortable just assuming that large numbers of homebuyers follow ad hoc extrapolative beliefs, we prefer assuming smaller deviations from rationality since these deviations must apply to such a large number of homebuyers.

One means of microfounding extrapolative beliefs is to assume that investors are unsure about the growth rate in the fundamental. We illustrate our point with a particularly simple model in which $R_t = R_{t-1} + g_R$. The correct pricing formula satisfies $P_t = R_t + \beta P_{t+1}$. The correct pricing formula is that $P_t = \frac{R_t}{1-\beta} + \frac{\beta E_t(g_R)}{(1-\beta)^2}$. At all points of t , individuals observe R_t , but they will not observe past values of R and will only observe the past two values of P .

At time zero, individuals know R_0 , but have no information about the growth rate and use their prior belief, $g_R = 0$, in their pricing. Hence, the price at time 0 equals $\frac{R_0}{1-\beta}$. At time one, new buyers emerge who observe R_1 and P_0 . If they are rational, then they can readily uncover the true value of g_R , which equals $R_1 - (1-\beta)P_0$. The new price equals $\frac{R_1}{(1-\beta)^2} - \frac{\beta P_0}{1-\beta}$, which is completely rational. The heavy weight given to R_1 reflects its double role in determining the flow of welfare associated with the house and also in determining the growth rate of that flow.

At time 2, if buyers are sophisticated, they can look at prices at time zero and time one and we now turn to the price at time 2. If time 2 buyers were sophisticated, $t+2$. If these buyers were sophisticated, they would look at prices at time zero and time one, and the correct pricing formula would be

$$P_2 = \frac{R_2}{1-\beta} + \beta(P_1 - P_0) = P_1 + (1-\beta)(P_1 - P_0). \quad (11.6)$$

There would be momentum but only because there is a persistent growth rate. The prices would not be growing up any faster than that fixed growth rate.

However, assume that these buyers were less sophisticated. They follow a rule-of-thumb assumption that prices at all times reflect the same assumption about growth rates, since growth rates are not changing. Since they do not observe the growth rate, they cannot directly infer the past levels of R but instead use the single change in P that they observe to infer a growth rate. They believe that $P_{t+1} - P_t$ equals $\frac{R_{t+1}-R_t}{1-\beta}$ or $\frac{g_R}{1-\beta}$. Hence, the price at time t will equal $\frac{R_t}{1-\beta} + \frac{\beta(P_{t-1}-P_{t-2})}{1-\beta}$. This implies

$$P_2 = \frac{R_2}{1-\beta} + \frac{\beta(P_1 - P_0)}{1-\beta} = P_1 + \left(1 - \beta + \frac{\beta^2}{1-\beta}\right)(P_1 - P_0). \quad (11.6')$$

The level of momentum has increased, perhaps dramatically, because the growth rate is being inferred from the change in prices, which themselves already included changing expectations about the growth rate.

In essence, these buyers are cursed in the phrase of [Eyster and Rabin \(2005\)](#), which means that individuals mistakenly believe that “other players’ actions depend less on their types than they actually do.” [Eyster and Rabin’s \(2005\)](#) idea is that it is difficult to fully infer the motives of others, and as a result, we don’t always make sense of market outcomes. For example, to avoid the winner’s curse, bidders must understand that if others bid less, then they must have worse signals about the value of the good being sold. Individuals will take bets when they shouldn’t because they don’t recognize that an individual who offers to bet with them is implicitly signaling some private information.

Glaeser and Nathanson develop a more sophisticated version of this approach in which growth rates are stochastic, and past prices are informative about the current rate of fundamental growth. The critical weakness again that delivers extrapolation, and

hence positive serial correlation, is that homebuyers believe that past price movements reflect changes in fundamentals, rather than changes in beliefs about the growth rate.

11.5.3 Cognitive limitations: natural expectations, spatial benchmarking, and rule-of-thumb spending

That particular model is part of a general class of models of cognitive limitations, many of which have the potential to add to our understanding of real estate fluctuations. [Shiller \(1999\)](#) reviewed many of the standard behavioral tics, which may influence housing markets. The intellectual challenge is that there are so many potential, plausible cognitive limitations that it is hard to naturally focus on any particular one. We will discuss several types of cognitive limitation here, and the extrapolation discussed above can also be seen as one example of cognitive limitation: the inability to think through how other people think.

[Fuster et al. \(2010\)](#) offered one particular form of cognitive limitation that they title natural expectations. Natural expectations models require agents to make predictions using only an excessively parsimonious model. For example, if the true R_t process was described as $R_t - R_{t-1} = B_1(R_{t-1} - R_{t-2}) - B_2(R_{t-2} - R_{t-3}) + \epsilon_t$, then the agents might attempt to estimate the regression by fitting only $R_t - R_{t-1} = \hat{B}_1(R_{t-1} - R_{t-2}) + \epsilon_t$. This obviously has the capacity to create mistakes, but it does naturally generate excess volatility or strong one-period price momentum.

If homebuyers have rational expectations, then applying the usual pricing formula ($P_t = R_t + \frac{1}{1+r}E(P_{t+1})$) implies

$$P_t = \frac{(1+r)^3 R_t - (1+r)((1+r)B_1 - rB_2)R_{t-1} - (1+r)^2 B_2 R_{t-2}}{r(1+r)(1+r-B_1) - rB_2}. \quad (11.7)$$

Homebuyers with natural expectations would set

$$P_t = \frac{(1+r)^2 R_t - (1+r)\hat{B}_1 R_{t-1}}{r(1+r-\hat{B}_1)}. \quad (11.7')$$

In [Table 11.2](#), we simulate the impact of these beliefs assuming that $B_1 = \hat{B}_1 = 0.9$, $B_2 = 0.8$, and that $B_1 = \hat{B}_1 = 0.5$, $B_2 = 0.3$. The interest rate is 0.04 and the standard deviation of the shock is \$1000. We do not believe that these parameter choices are realistic. They are instead chosen to illustrate what is necessary for natural expectations to deliver high degrees of momentum and excess volatility. The medium persistence case still has considerably more momentum than either income or rents. In that case, the volatility is slightly increased by moving from rational to natural expectations. There is considerably more mean reversion with natural expectations, because the buyers do not realize that positive shocks today will lead to negative shocks in two periods. However, this does not generate significant price momentum. Since even the natural expectations

Table 11.2 Fixed supply model

(1)	(2)	(3)	(4)	(5)
Horizon	Medium persistence		High persistence	
	Rational	Projection	Rational	Projection
Price change volatility				
1 year	32,600	52,800	29,800	349,000
3 years	56,700	86,000	51,300	649,000
5 years	72,900	94,000	65,000	434,600
Price change serial correlation				
1 year	0.01	0.05	0.01	0.35
3 years	-0.01	-0.34	-0.04	-0.85
5 years	0.00	0.20	0.00	-0.20

agents recognize that a shock today will become a shock tomorrow, this positive effect is built immediately into prices and this means that there is little extra momentum created by this cognitive limitation.

In the truly extreme case of massive short-run persistence and massive medium-run mean reversion, we do indeed get momentum in prices with natural expectations but not with rational expectations. The rational expectations agents understand that a positive shock today will be offset soon and, as a result, prices move far less. The natural expectations agents have much more severe price movements and significant momentum. This momentum turns into mean reversion when the true nature of the process inserts itself. Of course, these natural expectations agents are making mistakes on a massive scale with some degree of regularity, which may be difficult for many economists to accept. We find this exercise interesting but believe that the assumptions needed for natural expectations to fit the housing price data are just too extreme.

One particularly important use of natural expectations, however, is that they may explain why homebuyers so often appear to miss the power of supply to bring prices back down to earth. If it always takes time to build new housing units, then making predictions based on short time periods will always mean missing the power of supply. This will mean that natural expectations will typically lead to a demand-side analysis only and lead buyers to fail to predict that supply will eventually cause prices to converge.

To illustrate this point, we assume that $R_t = \theta_t - \alpha N_t$, where θ_t is the exogenous demand shock and N_t reflects the supply of new housing. We assume that $\theta_{t+1} - \theta_t = \delta + \rho(\theta_t - \theta_{t-1} - \delta) + \varepsilon_{t+1}$, so the true process is indeed a one-period moving average. We allow a two-period production process $E_t(P_{t+2}) = c_0 + c_1 I_{t+1}$ and $N_{t+1} = N_t + I_t$. This means that the production decision that determines the stock at $t+2$ is made with the information available as of time t .

Table 11.3 Variable supply with low construction costs

(1)	(2)	(3)	(4)	(5)	(6)
Horizon	Full rational	$\bar{I} = I_t$		$\bar{I} = I_{t-1}$	
		Semi	Rational	Semi	Rational
Price change volatility					
1 year	6000	50,400	51,300	51,600	52,500
3 years	9700	50,500	51,300	73,000	74,300
5 years	11,300	50,700	51,300	73,000	74,300
Construction correlation					
1 year	5500	2000	2000	2000	2000
3 years	14,600	5900	6000	5900	6000
5 years	22,300	9600	9800	9600	9800
Construction serial correlation					
1 year	0.75	0.96	0.96	0.96	0.96
3 years	0.55	0.92	0.92	0.92	0.92
5 years	0.41	0.88	0.87	0.88	0.87

We consider three possibilities. First, it is possible that both the homebuyers and the homebuilders are completely rational. This is the case shown in the first column of [Table 11.3](#). In this case, there is modest volatility, significant mean reversion, and little price momentum. Buyers anticipate all the future changes, but it remains true that a new building causes initial shocks to disappear over time.

In all cases, we allow the homeowners to correctly understand the dynamics of demand but not to fully understand the dynamics of supply. In both of our semirational cases, we assume that homebuyers believe that supply is fixed. In columns (2) and (3), buyers believe that supply is fixed at the current rate of supply. In columns (4) and (5), buyers believe that supply will be fixed at the level supplied next period. We also have two possibilities for the degree of rationality about homebuilders. In columns (2) and (4), homebuilders choose I_{t+1} , assuming that $I_{t+1} = I_{t+2}$. In columns (3) and (5), builders are completely rational.

In all cases, these near-rational assumptions fail to deliver any price momentum. The failure to anticipate supply responses just does not deliver a reason for price growth to follow price growth. There is momentum in the demand fundamentals, but the natural expectations buyers build that into their period t prices. However, the buyers' limited ability to anticipate supply does exacerbate price volatility and price mean reversion. The price volatility is higher because they do not anticipate the fact that rents will be declining over time as new supply enters into the market. Mean reversion is higher because prices rise more initially but then come back down to earth quickly.

Somewhat surprisingly to us, this semirationality reduces the volatility of construction changes. As prices move around a great deal, construction moves less. A second less studied form of cognitive limitation is spatial benchmarking. This rule of thumb takes the spatial equilibrium logic of the Alonso–Muth–Mills model and the Rosen–Roback model and applies it to prices, rather than rents. This type of logic was used historically to convince investors in the wisdom of Los Angeles real estate prices during the boom of the 1880s, and it is compatible with the [Ferreira and Gyourko \(2012\)](#) evidence on the spatial spread of the boom during the last decade.

This logic will be particularly problematic if there are supply differences across areas. To consider an extreme example, assume that in city A (Los Angeles), the supply is fixed and $R_{t+j} = e^{gj} R_t$, so the benefits are deterministically growing. Applying the pricing formula $P_t = \int_{j=0}^{\infty} e^{-rj} R_{t+j} dj$ implies that prices in city A should equal $P_t = \frac{R_t}{r-g}$.

Assume that at a point in time, city B yields benefits $R_t - \delta$. Naive spatial extrapolation would then imply that the price in city B should equal $\frac{R_t - \delta}{r-g}$ and this would indeed be rational if prices in city B were also increasing at the rate g . One reason why the formula might be radically wrong is that the growth rates differ in the two areas. In that case, static comparisons would lead to incorrect pricing.

A less obvious source of error can come from ignoring supply conditions across the two areas. To focus on that possibility, we assume that $g=0$ and R_t is fixed at R_A in city A . The fixed flow of utility reflects the assumption that supply in city A is also fixed. Prices in city A should equal $\frac{R_A}{r}$.

In city B , $R_t = \theta_B - \alpha N_t$, so the flow utility is falling with the number of people living in the city, perhaps because of congestion. At time zero, $\theta_B - \alpha N_0 = R_A$, so initially the two places yield comparable returns. Hence, naive spatial extrapolation implies that the price in both cities will equal $\frac{R_A}{r}$. This will differ from the rational price because individuals should expect the impact of new supply. The cost of supplying new homes is increasing linearly in the number of new homes, so that the price must equal $c_0 + c_1 \dot{N}$, whenever new homes are being built. We assume $\frac{R_A}{1-\beta} > c_0$ to ensure that new construction always occurs.

The rational pricing and growth equations in city B satisfy

$$\dot{N} = \left(\frac{\theta_B - \alpha N_0}{r} - c_0 \right) \frac{r(r + c_1)}{c_1 r(r + c_1) + \alpha} e^{-c_1 t} \quad (11.8)$$

and

$$P = c_0 + \left(\frac{\theta_B - \alpha N_0}{r} - c_0 \right) \frac{c_1 r(r + c_1)}{c_1 r(r + c_1) + \alpha} e^{-c_1 t}. \quad (11.9)$$

Even at time zero, the rational price in city B will equal $\frac{\alpha c_0}{c_1 r(r + c_1) + \alpha} + \left(\frac{\theta_B - \alpha N_0}{r} \right) \frac{c_1 r(r + c_1)}{c_1 r(r + c_1) + \alpha}$, which must be less than $\frac{R_A}{r}$, the naive spatial expectations price. Naive buyers in city B are comparing their city with city A , perhaps like buyers in Las Vegas in 2005 comparing

their city with Los Angeles. They see comparable flow utility in the two cities, which suggests to them, naively, that the prices should be comparable. Yet, sophisticated buyers understand that city *B* will add housing over time, which will cause flow utility to fall, and that city *B* housing should therefore cost less than city *A*.

If individuals in city *B* persist in using the pricing formula that is appropriate for city *A*, so that prices equal $\frac{\theta_B - \alpha N_t}{r}$, then $\dot{N} = \frac{1}{c_1} \left(\frac{\theta_B - \alpha N_t}{r} - c_0 \right)$, which will initially be larger than under rational pricing. This will lead to overbuilding in the short run, and eventually, prices will lie below their level under rational pricing. The rents in the two cities will start at the same level, but they will soon differ because of excess supply, and initially, prices should reflect this expected convergence.

A third way in which prices may diverge from rationality is that buyers follow other prescribed rules of thumb. Some options that have been discussed are always spending as much as they can afford, given current interest rates. If the marginal homebuyer has an income of *Y* dollars and is able to get a no-down payment mortgage, then the maximum willingness to pay is some fixed fraction times *Y* divided by the interest rate. This rule of thumb suggests a high elasticity of price with respect to interest rates but one that is no different than in the standard model. It also suggests that the price of housing will be decoupled from the benefits of housing, which does run counter to centuries of economic thinking. Money illusion can also contribute to “housing frenzies” if buyers overestimate the future costs of real interest payments when inflation is high and will then bid more for housing as inflation drops (Brunnermeier and Julliard, 2008).

One added possibility is that individuals hop from one type of belief formation to another. For example, they may sometimes be rational but become adaptive during periods of sustained growth. Alternatively, they may just ignore future price appreciation much of the time during periods when such appreciation seems unlikely or is just salient. During those periods, prices track rents, but if an event makes the prospect housing price appreciation salient, then buyers start making potentially biased forecasts about future housing price growth.

11.5.4 Social learning and the entrepreneurs of error

The previous section focused on learning from past price movements, but there are many other influences that shape individuals’ beliefs about housing prices. Perhaps, the most pervasive and important source of information that humans rely upon is the statements of people around us. Most of the time, these social influences are relatively benign, but in some cases, these influences may reflect private motives at odds with the individuals’ own best motives.

The early literature on manias emphasized the rational causes of imitation. Froot et al. (1992), for example, emphasized the strong incentives of investors with short horizons to focus on the same sources of information. Bikhchandani et al. (1998) and Banerjee (1992)

both presented models of information cascades, where individuals rationally imitate one another. DeCoster and Strange (2012) applied this logic to developers, who imitate one another because they assume that their peers have made their decisions based on valuable information about the state of the world. The result can be a glut of overbuilding.

These forces will only become more powerful if the urge to imitate exceeds the purely rational. One natural version of this is to again assume a type of “cursed” behavior, where individuals underestimate the social causes of the behavior of others. In this case, each new buyer infers that the mass of preceding buyers is acting on private information, rather than just following the leader. In this case, each new buyer believes that the actions of the herd contain an extraordinary rich amount of information, whereas in reality, the mob might just be following the leader of a single person. This type of incorrect inference will tend to make herd behavior extremely powerful and manias extremely common.

While DeCoster and Strange emphasized builders’ decisions, this same logic could relate to buying homes in a particular locale, such as Las Vegas in 2005. According to this view, the large number of Las Vegas buyers provides evidence to new buyers that Las Vegas is an excellent investment. This logic then encourages an even larger rush of buyers.

The literature has expanded beyond imitating actions to following advice. One notable paper in this genre is Hong et al. (2008), which argues that interested advisors may play an important role in encouraging the role of bubbles. According to this view, there are individuals who have an interest in selling stock or real estate. These advisors provide misleading information to buyers who then act on these incorrect messages. Real estate agents do have the motive to encourage buyers to bid and are typically given plenty of time in which to make that case.

Naturally, these models assume a degree of irrationality—individuals still listen to advisors who are patently self-interested. Yet, this attention to the opinions of others may itself be an entirely sensible rule of thumb. Most of the time, advice is given disinterestedly (i.e., one’s spouse or mother advises that you wear a coat because it is raining), and it is best not to waste too much effort trying to understand the motives behind the advice. Perhaps, we follow the advice that we are given because that is a relatively sensible strategy most of the time.

One attractive aspect of the “entrepreneurs of error” approach is that it appears to offer more testable implications than the simple herding view. Herding models, like agglomeration models in urban economics, yield the prediction that individuals will act similarly. Yet, it remains unclear whether the correlation across individuals reflects herding (or agglomeration) or just some omitted factor that acts on everyone simultaneously. By contrast, the “entrepreneurs of error” approach offers the strong prediction that herds will move in the direction implied by the error suppliers with strong incentives and persuasive talents.

We are not ruling out any number of possible models with cognitive limitations, but this research agenda is sufficiently early that we suspect that concentrating on a small

number of alternatives to complete rationality is sensible. We suspect that there is an agenda around extrapolative beliefs involving theoretical research, normal empirics, and lab work that will be highly productive. A better understanding of why the implications of elastic supply seem to be so often ignored also seems relevant. Finally, for exploring some elements of housing institutions or policies, it will remain sensible to take the easy if hard to defend approach of just assuming exogenous beliefs.

11.5.5 Directions for future research

We are at the dawn of research on near-rational models of housing markets and housing bubbles. On the theoretical front, we are particularly enthusiastic about models with simple deviations from rationality that can produce the stylized facts discussed in [Section 11.3](#). Ideally, we would be able to focus on a single form of near-rationality, rather than having an explanation for each anomaly. We think that microfounded extrapolation models and social learning models are particularly likely to be worth further investigation.

There is also a need for a broader set of empirical tools in this area. Many behavioral quirks have been investigated in the lab, but lab conditions are far from the conditions experienced when families are shopping with real estate agents. Field experiments would seem like a priority here. More generally, good behavioral models will yield new testable implications, and those implications will enable us to judge their further value.

There is also a specific need for normative analysis with near-rational homebuyers. What should the optimal policies towards the home mortgage interest deduction be when homebuyers are overoptimistic or trend chasing? How do behavioral idiosyncrasies relate to land use controls? Any normative analysis should also presumably pay some attention to the possible near-rationality on the part of regulators, legislators, and voters.

11.6. PUBLIC POLICY AND BUBBLES

Real estate bubbles relate to public policy both positively and normatively. On the positive side, many observers have argued that government policies, including the low interest rates, the Community Reinvestment Act, and support for Fannie Mae and Freddie Mac, helped to cause the bubble and its burst (e.g., [Wallison, 2009](#)). We have already argued that it is hard to see any simple connection between easy credit and the housing bubble, which leads us to be cautious about accepting the view that such policies obviously caused the boom. It remains, of course, possible that these policies exacerbated the bubble and its financial implications, but it seems hard to blame the government for the fluctuation. Moreover, the history of real estate bubbles suggests that they have often occurred when government intervention is minimal ([Glaeser, 2013](#)).

We turn to the normative, public policy implications of housing bubbles for different parts of the public sector. Most obviously, macroeconomic institutions, such as the Federal Reserve Board, debate whether to engage in policies explicitly aimed at reducing

the volatility of asset bubbles, including real estate bubbles. Bank and credit regulators have oversight over institutions that are deeply impacted by real estate fluctuations. How does the existence of real estate bubbles impact optimal banking regulations? The federal government engages in a series of housing market policies, including the home mortgage interest deduction. Should these policies be changed in light of recent real estate volatility? Finally, local land use regulations are largely responsible for shaping housing supply. These regulations also interact with housing bubbles (Glaeser et al., 2008).

We have no intention of resolving these policy issues now, but we note them primarily as enduringly important topics. The Federal Reserve Board has historically abstained from taking steps to deflate asset bubbles. For example, in 2005, Janet Yellen articulated the current orthodoxy that monetary policy should not be used to deflate a housing bubble. She asked “if the bubble were to deflate on its own, would the effect on the economy be exceedingly large?” and refuted that claim. More importantly, she argued that monetary policy is not “the best tool to use to deflate a house-price bubble.”

History has not been kind to her first argument against intervention. The real estate bust did have widespread adverse consequences and the threat of future housing busts seems quite real. To be fair, many housing economists (including at least one of us) were at least as mistaken as she was. Still, after learning the Great Recession, future policy makers should never be so confident that a housing downturn won’t have serious consequences.

But what awareness of the risks means for macroeconomic policy making is far less clear. Yellen’s point that monetary policy is not a good way to “deflate a house-price bubble” is just as tenable today as it was in 2005. Moreover, since housing price booms can reflect real forces, as well as bubbles, it may be foolish to constantly attempt to run counter to rising prices. The policy conclusion is uncertain, but that provides a far clear implication that more research is needed on macroeconomic stabilization policy when real estate volatility is large.

The volatility of real estate prices also impacts financial market regulation. The Lehman Brothers’ bankruptcy was closely connected with its exposure to real estate-related subprime mortgage risk. Indeed, the securitization of mortgages has been blamed not only for helping to create the bubble but also for ensuring that the pain of the downturn is more widely experienced. Yet, the spreading of that risk may have reduced the adverse consequences of the bubble for the banking system itself, since fewer mortgages were being held directly on the books of lending institutions.

The obvious implication of centuries of real estate booms and busts is that real estate is not a riskless asset. Regular mean reversion means that high prices today may well mean low prices tomorrow. Presumably, these facts should inform banking regulation if the goal of such regulation is to reduce the risk of financial distress within the sector.

One suggested reform is that regulators should anticipate mean reversion when assessing asset values for capital requirements. If prices have risen by 75% over the past 5 years, then historical experience suggests that a 25% drop over the next 5 years is not unreasonable. One proposal is to value real estate-related capital based on its future expected value.

Yet, there are many reasons to be cautious about changes of this kind. Real estate is not the only asset that displays mean reversion (Cutler et al., 1991), but if real estate is the only asset that is subject to such treatment, then this may distort the movement of capital. Whatever formula is used to assess long-run value will surely be subject to gaming by lending institutions and political influence by policy makers seeking their own pet objectives. Again, the only conclusion that can be definitively drawn now is the need for further investigation.

The federal government has explicit policies that promote homeownership typically by subsidizing lending. The home mortgage interest deduction implicitly subsidizes home borrowing. While borrowing for business investments may also be deductible, typically, the returns to those investments are taxed. By contrast, the government does not tax the implicit rental income earned by a homeowner. The government-sponsored enterprises, Freddie Mac and Fannie Mae, as well as the Federal Housing Administration, have also all encouraged home borrowing by providing a guarantee against default.

The presence of real estate bubbles matters deeply for the wisdom of encouraging leveraged bets on real estate through public policy. If real estate was a safe, boring asset that rarely experienced major fluctuations, then encouraging homeownership could be seen as a safe means of encouraging asset accumulation. However, high levels of volatility mean that government policies that support leveraged borrowing can have the unfortunate impact of creating a class of homeowners who are massively indebted because they borrowed to buy housing that has lost its value.

To a certain extent, this downside risk is offset by the fact that the cost of living for these homeowners has dropped. Owning itself is something of a hedge since we are all born short housing (Sinai and Souleles, 2008). Yet, there is also a covariance between local housing prices and local labor markets, which helps explain why the downturn was associated with so many foreclosures. Those foreclosures provide a tangible example of the risks associated with encouraging leveraged REITs particularly for lower income Americans.

The supporters of pro-home borrowing policies will often point out the large share of housing in the portfolios of many Americans, as if this proves that subsidizing homeownership is the natural path towards encouraging asset accumulation. Yet, since many of these policies make it easier to borrow with low down payments, they also reduce the incentive to save before buying. There is a severe need for a more serious literature about the portfolio implications for ordinary Americans of encouraging home borrowing, especially in light of significant house value volatility.

The final relevant policy area relating to real estate bubbles concerns the land use policies of the local government. Housing supply is determined, at least in part, by regulations at the local level. Housing supply then in turn influences the nature and duration of real estate bubbles. The first-order correlation is that these events do appear to be more extreme in more restricted areas, both in places that currently cannot build (Glaeser et al., 2008) and in places where investors anticipate future regulatory constraints that make supply difficult (Nathanson and Zwick, 2013).

Does it therefore follow that in a world with real estate bubbles, it is even more valuable to reduce the barriers to new building? Not necessarily. Even if we were confident that fewer restrictions on buildings might make bubbles less common or less extreme, we could not conclude that reducing land use restrictions would reduce the social costs of bubbles. For one of those costs is overbuilding, and overbuilding will be more severe in places where land use is more restricted. It seems doubtful, for example, that the price boom between 2001 and 2006 caused any serious overbuilding in San Francisco or the suburbs of Boston, because so little new housing was built and prices remained significantly above construction costs even after the bust. Their restrictions may have caused price swings to be more severe, but they also limited the hangover from excess building supply. The social cost of overbuilding will be most severe in areas in which supply is elastic and bubbles still occur.

Overall, this policy section has provided no clear policy directives and that is precisely the point. The policy implications of real estate bubbles are far from clear. Economists have not spent all that much time researching this issue, partially because of our unwillingness to accept the existence of bubbles. Yet, the hair-raising events of 2000–2012 make it plain that real estate can experience enormous convulsions, whether we choose to use the word “bubble” or not. Surely, economists must press on to deliver a better apparatus for understanding the consequences of different housing and banking policies in a world where real estate can be very volatile.

The need for policy-related research is enormous. We certainly believe that the positive questions discussed above need to be answered before we can be confident about the bases of public policy, but policy making must often proceed in environments of ignorance. One of the biggest questions is whether regulatory requirements, such as higher reserve ratios, can effectively protect the financial system against future housing busts. The question remains unresolved. Some experts have even argued that higher interest rates are more likely to be effective than regulation because they can act system-wide, impacting the shadow banking sector, for example, in ways that regulation cannot.

If societies contemplate changing the incentives that subsidize home lending, other questions become particularly important. Does the current regime of subsidizing home-buying actually lead to added asset accumulation, or does it reduce the incentive to save for down payments? What does subsidized home borrowing do to the risk profile of normal households?

11.7. CONCLUSION

This chapter has argued that real estate experiences impressive booms and busts, which can reasonably be referred to as bubbles. Generally, prices move too much to be fully explained by changes in either rents or observable fundamentals. Housing prices display substantial momentum at high frequencies, and they mean revert at lower frequencies. These general features were greatly exacerbated during the great boom and bust of the 2000–2012 period. Moreover, real estate convulsions have appeared regularly throughout the United States and world history often with dire consequences.

The economics of real estate bubbles is still in its infancy, for until 2005, the dominant economic view was that such bubbles do not exist. This orthodoxy paralleled the general assumption in financial economics, at least until the dot-com bust of 2000. It seems silly now to believe that housing price changes are orderly and driven entirely by obvious changes in fundamentals operating through a standard model.

Moving ahead, there are two broad classes of models that have and will continue to shed light on these great housing convulsions. The first class essentially attempts to explain these housing market features with essentially rational actors. We have discussed two variants of these models. First, several papers examine housing dynamics through the lens of a search model. These models can generate substantially more momentum than standard models, because it takes time for shocks to work their way through the system. To date, these models generally do not deliver large amounts of excess volatility, just momentum and mean reversion, but those are themselves significant contributions, and it remains possible that future models of learning and search will also generate significant excess volatility, although there are reasons to doubt that this will occur.

A second form of rational model stresses that if agency problems lead to interest rates that charge too little for default risk, then rational bubbles can occur without violating any transversality conditions. These models still tend to predict that prices will become enormously high with positive probability. This implication is difficult to square with elastic housing supply, which suggests that these events are far more plausible in highly constrained areas. These rational bubble models do not imply that cheap credit will always cause a real estate bubble, but they do suggest that cheap credit is a necessary condition.

The second class of models drops the assumption of perfect rationality. We have discussed several different types of these models, including models in which beliefs were just assumed to be fixed and heterogeneous. While that assumption is useful for highlighting cross-sectional aspects of the housing markets, we suspect that productive veins of behavioral real estate research going forward will focus on extrapolative beliefs and simple forms of limited cognition. We sketched a path towards grounding extrapolation as the result of cognitive limitations of the form discussed by [Eyster and Rabin \(2005\)](#). It also seems possible that cognitive limitations may lie behind the buyers' apparent tendency to repeatedly ignore the power of housing supply to determine prices.

Understanding the causes of real estate bubbles seems particularly critical, because these events clearly have large social consequences. We cannot plausibly hope that these asset price fluctuations will vanish, but it is at least possible that we can reduce their costs through better public policy making. Unfortunately, we are still far from having enough knowledge to confidently recommend any particular policy actions.

ACKNOWLEDGMENT

Glaeser thanks the Taubman Center for State and Local Government for the financial support. William Strange (the editor) provided much guidance, and Rajiv Sethi and William Strange both provided excellent comments. This chapter is written for the Handbook of Urban and Regional Economics.

REFERENCES

- Abel, J.R., Deitz, R., 2010. Bypassing the bust: the stability of upstate New York's housing markets during the recession. *Fed. Res. Bank N.Y. Curr. Iss. Econ. Finance* 16 (3), 1–9.
- Abreu, D., Brunnermeier, M.K., 2003. Bubbles and crashes. *Econometrica* 71 (1), 173–204.
- Adelino, M., Gerardi, K., Willen, P., 2010. What explains differences in foreclosure rates? A response to Piskorski, Seru, and Vig: Federal Reserve Bank of Atlanta Working paper 2010–8, Federal Reserve Bank of Atlanta, Atlanta, GA.
- Ambrose, B.W., Eichholtz, P., Lindenthal, T., 2013. House prices and fundamentals: 355 years of evidence. *J. Money, Credit, Bank.* 45 (2–3), 477–491.
- Banerjee, A.V., 1992. A simple model of herd behavior. *Q. J. Econ.* 107 (3), 797–817.
- Barberis, N., Greenwood, R., Jin, L., Shleifer, A., 2015. X-CAPM: an extrapolative capital asset pricing model. *J. Financ. Econ.* 115 (1), 1–24.
- Becker, G.S., 1962. Irrational behavior and economic theory. *J. Polit. Econ.* 70 (1), 1–13.
- Bikhchandani, S., Hirshleifer, D., Welch, I., 1998. Learning from the behavior of others: conformity, fads, and informational cascades. *J. Econ. Perspect.* 12 (3), 151–170.
- Brunnermeier, M.K., Julliard, C., 2008. Money illusion and housing frenzies. *Rev. Financ. Stud.* 21 (1), 135–180.
- Caldera, A., Johansson, A., 2013. The price responsiveness of housing supply in OECD countries. *J. Hous. Econ.* 22 (3), 231–249.
- Campbell, J.Y., Giglio, S., Pathak, P., 2011. Forced sales and house prices. *Am. Econ. Rev.* 101 (5), 2108–2131.
- Case, K.E., Shiller, R.J., 1989. The efficiency of the market for single family homes. *Am. Econ. Rev.* 79 (1), 125–137.
- Case, K.E., Shiller, R.J., 2003. Is there a bubble in the housing market? *Brook. Pap. Econ. Act.* 2, 299–362.
- Case, K.E., Quigley, J.M., Shiller, R.J., 2005. Comparing wealth effects: the stock market versus the housing market. *Adv. Macroecon.* 5 (1), 1–32.
- Case, K.E., Shiller, R.J., Thompson, A., 2012. What have they been thinking? Homebuyer behavior in hot and cold markets. *Brook. Pap. Econ. Act.* 2, 265–315.
- Clapp, J.M., Tirtiroglu, D., 1994. Positive feedback trading and diffusion of asset price changes: evidence from housing transactions. *J. Econ. Behav. Organ.* 24 (3), 337–355.
- Clapp, J.M., Dolde, W., Tirtiroglu, D., 1995. Imperfect information and investor inferences from housing price dynamics. *Real Estate Econ.* 23 (3), 239–269.
- Cox, J.C., Ingersoll Jr., J.E., Ross, S.A., 1985. A theory of the term structure of interest rates. *Econometrica* 53 (2), 385–407.
- Cutler, D.M., Poterba, J.M., Summers, L.H., 1991. Speculative dynamics. *Rev. Econ. Stud.* 58 (3), 529–546.

- Davidoff, T., 2013. Supply elasticity and the housing cycle of the 2000s. *Real Estate Econ.* 41 (4), 793–813.
- DeCoster, G.P., Strange, W.C., 2012. Developers, herding, and overbuilding. *J. Real Estate Finance* 44 (1–2), 7–35.
- DeLong, J.B., Shleifer, A., Summers, L.H., Waldmann, R.J., 1990. Positive feedback investment strategies and destabilizing rational speculation. *J. Financ.* 45 (2), 379–395.
- Demsetz, R.S., Saidenberg, M.R., Strahan, P.E., 1997. Agency problems and risk taking at banks: Federal Reserve Bank of New York Research Paper 9709, Federal Bank of New York, New York, NY.
- Diamond, D.W., Rajan, R., 2009. The credit crisis: conjectures about causes and remedies. *Am. Econ. Rev.* 99 (2), 606–610.
- Diba, B.T., Grossman, H.I., 1988. The theory of rational bubbles in stock prices. *Econ. J.* 98 (392), 746–754.
- Eyster, E., Rabin, M., 2005. Cursed equilibrium. *Econometrica* 73 (5), 1623–1672.
- Fama, E.F., 1965. The behavior of stock-market prices. *J. Bus.* 38 (1), 34–105.
- Ferreira, F., Gyourko, J., 2012. Heterogeneity in neighborhood-level price growth in the United States, 1993–2009. *Am. Econ. Rev.* 102 (3), 134–140.
- Ferreira, F., Gyourko, J., Tracy, J., 2010. Housing busts and household mobility. *J. Urban Econ.* 68 (1), 34–45.
- Fisher, L.M., Lambie-Hanson, L., Willen, P.S., 2013. The role of proximity in foreclosure externalities: evidence from condominiums: The Federal Reserve Bank of Boston Public Policy Discussion Paper No. 13–2, Federal Reserve Bank of Boston, Boston, MA.
- Froot, K.A., Scharfstein, D.S., Stein, J.C., 1992. Herd on the street: informational inefficiencies in a market with short-term speculation. *J. Financ.* 47 (4), 1461–1484.
- Fuster, A., Laibson, D., Mendel, B., 2010. Natural expectations and macroeconomic fluctuations. *J. Econ. Perspect.* 24 (4), 67–84.
- Gao, Z., 2014. Housing boom and bust with elastic supplies: Job Market Paper. Princeton University, Princeton, NJ.
- Gelain, P., Lansing, K.J., 2014. House prices, expectations, and time-varying fundamentals. *J. Empir. Finance* 29, 3–25.
- Genesove, D., Han, L., 2012. Search and matching in the housing market. *J. Urban Econ.* 72 (1), 31–45.
- Genesove, D., Mayer, C., 2001. Loss aversion and seller behavior: evidence from the housing market. *Q. J. Econ.* 116 (4), 1233–1260.
- Giglio, S., Maggiori, M., Stroebel, J., 2014. No-bubble condition: model-free tests in housing markets: Working paper #20154, National Bureau of Economic Research, Cambridge, MA.
- Glaeser, E.L., 2013. A nation of gamblers: real estate speculation and american history. *Am. Econ. Rev.* 103 (3), 1–42.
- Glaeser, E., Gyourko, J., 2009. Rethinking Federal Housing Policy. American Enterprise Institute, Washington DC.
- Glaeser, E.L., Gyourko, J., Saiz, A., 2008. Housing supply and housing bubbles. *J. Urban Econ.* 64 (2), 198–217.
- Glaeser, E., Gottlieb, J.D., Gyourko, J., 2013. Can cheap credit explain the housing boom? In: Glaeser, E.L., Sinai, T. (Eds.), *Housing and the Financial Crisis*. University of Chicago Press, Chicago, IL, pp. 301–359.
- Glaeser, E.L., Gyourko, J., Morales, E., Nathanson, C.G., 2014. Housing dynamics: an urban approach. *J. Urban Econ.* 81 (2014), 45–56.
- Goodman, J., 2005. Housing Situation and Time Use. Advisors, Hartrey.
- Green, R.K., 2008. Imperfect information and the housing finance crisis: a descriptive overview. *J. Hous. Econ.* 17 (4), 262–271.
- Greenwood, R., Hanson, S.G., 2013. Issuer quality and corporate bond returns. *Rev. Financ. Stud.* 26 (6), 1483–1525.
- Guren, A., 2014. The causes and consequences of house price momentum: Job Market Paper. Harvard University, Cambridge, MA.
- Guren, A., McQuade, T., 2014. How Do Foreclosures Exacerbate Housing Downturns? Boston University, Boston, MA.

- Hall, R.E., 2003. *Deal Engines: The Science of Auctions, Stock Markets, and E-Markets*. W. W. Norton & Company, New York, NY.
- Haughwout, A., Lee, D., Tracy, J.S., Wilber, V.d.K., 2011. Real estate investors, the leverage cycle, and the housing market crisis: Federal Reserve Bank of New York Staff Report No. 514, Federal Reserve Bank of New York, New York, NY.
- Head, A., Lloyd-Ellis, H., Sun, H., 2014. Search, liquidity, and the dynamics of house prices and construction. *Am. Econ. Rev.* 104 (4), 1172–1210.
- Himmelberg, C., Mayer, C., Sinai, T., 2005. Assessing high house prices: bubbles, fundamentals, and misperceptions. *J. Econ. Perspect.* 19 (4), 67–92.
- Hoberg, G., Phillips, G., 2010. Real and financial industry booms and busts. *J. Financ.* 65 (1), 45–86.
- Hong, H., Scheinkman, J., Xiong, W., 2006. Asset float and speculative bubbles. *J. Financ.* 61 (3), 1073–1117.
- Hong, H., Scheinkman, J., Xiong, W., 2008. Advisors and asset prices: a model of the origins of bubbles. *J. Financ. Econ.* 89 (2), 268–287.
- Kindleberger, C.P., 1978. *Manias, Panics, and Crashes: A History of Financial Crises*. Basic Books, New York, NY.
- Kivedal, B.K., 2013. Testing for rational bubbles in the US housing market. *J. Macroecon.* 38 (B), 369–381.
- Krainer, J., 2001. A theory of liquidity in residential real estate markets. *J. Urban Econ.* 49 (1), 32–53.
- Krainer, J., LeRoy, S.F., Munpyung, O., 2009. Mortgage default and mortgage valuation: Federal Reserve Bank of San Francisco Working paper No. 2009–20, Federal Reserve Bank of San Francisco, San Francisco, CA.
- Levitt, S.D., Syverson, C., 2008. Market distortions when agents are better informed: the value of information in real estate transactions. *Rev. Econ. Stat.* 90 (4), 599–611.
- Malinvaud, E., 1953. Capital accumulation and efficient allocation of resources. *Econometrica* 21 (2), 233–268.
- Mian, A., Sufi, A., 2009. The consequences of mortgage credit expansion: evidence from the US mortgage default crisis. *Q. J. Econ.* 124 (4), 1449–1496.
- Mian, A., Sufi, A., 2010. Household leverage and the recession of 2007–09. *IMF Econ. Rev.* 58 (1), 74–117.
- Mian, A., Sufi, A., 2014. *House of Debt: How They (and You) Caused the Great Recession, and How We Can Prevent It from Happening Again*. University Of Chicago Press, Chicago, IL.
- Mian, A., Rao, K., Sufi, A., 2013. Household balance sheets, consumption, and the economic slump. *Q. J. Econ.* 128 (4), 1687–1726.
- Mian, A., Sufi, A., Trebbi, F., 2014. Foreclosures, house prices, and the real economy. *J. Financ.* Forthcoming. <http://scholar.princeton.edu/atif/publication>.
- Nathanson, C.G., Zwick, E., 2013. Arrested development: theory and evidence of supply-side speculation in the housing market: Job Market Paper. Harvard University, Cambridge, MA.
- Novy-Marx, R., 2009. Hot and cold markets. *Real Estate Econ.* 37 (1), 1–22.
- Palmer, C., 2013. Why did so many subprime borrowers default during the crisis: loose credit or plummeting prices?: Job Market Paper. MIT, Cambridge, MA.
- Pavlov, A., Wachter, S.M., 2006. The inevitability of marketwide underpricing of mortgage default risk. *Real Estate Econ.* 34 (4), 479–496.
- Piazzesi, M., Schneider, M., 2009. Momentum traders in the housing market: survey evidence and a search model. *Am. Econ. Rev.* 99 (2), 406–411.
- Piskorski, T., Seru, A., Vig, V., 2010. Securitization and distressed loan renegotiation: evidence from the subprime mortgage crisis. *J. Financ. Econ.* 97 (3), 369–397.
- Poterba, J., 1984. Tax subsidies to owner-occupied housing: an asset market approach. *Q. J. Econ.* 99 (4), 729–745.
- Purnanandam, A., 2011. Originate-to-distribute model and the subprime mortgage crisis. *Rev. Financ. Stud.* 24 (6), 1881–1915.
- Roback, J., 1982. Wages, rents, and the quality of life. *J. Polit. Econ.* 90 (4), 1257–1278.
- Rosen, S., 1979. Wage-based indexes of urban quality of life. In: Mieszkowski, P., Straszheim, M. (Eds.), *Current Issues in Urban Economics*. Johns Hopkins University Press, Baltimore, MD.
- Santos, M.S., Woodford, M., 1997. Rational asset pricing bubbles. *Econometrica* 65 (1), 19–57.

- Scheinkman, J.A., Xiong, W., 2003. Overconfidence and speculative bubbles. *J. Polit. Econ.* 111 (6), 1183–1220.
- Shiller, R.J., 1981. Do stock prices move too much to be justified by subsequent changes in dividends? *Am. Econ. Rev.* 71 (3), 421–436.
- Shiller, R.J., 1999. Chapter 20: human behavior and the efficiency of the financial system. In: Taylor, J.B., Woodford, M. (Eds.), *Handbook of Macroeconomics*, vol. 1(1). Elsevier, Amsterdam, The Netherlands, pp. 1305–1340.
- Shiller, R.J., 2005. *Irrational Exuberance*. Random House LLC, New York, NY.
- Shleifer, A., Vishny, R.W., 1997. The limits of arbitrage. *J. Financ.* 52 (1), 35–55.
- Sinai, T., Souleles, N., 2008. Chapter 4: net worth and housing equity in retirement. In: Ameriks, J., Mitchell, O.S. (Eds.), *Recalibrating Retirement Spending and Saving*. Oxford University Press, Oxford, pp. 46–80.
- Sockin, M., Xiong, W., 2014. Learning about the neighborhood: a model of housing cycles: Princeton University Working paper, Princeton University, Princeton, NJ.
- Stein, J.C., 1995. Prices and trading volume in the housing market: a model with down-payment effects. *Q. J. Econ.* 110 (2), 379–406.
- Wallison, P.J., 2009. The true origins of this financial crisis. *Am. Spect.* <http://spectator.org/articles/42211/true-origins-financial-crisis> (accessed 13 August 2014).
- Wheaton, W.C., 1990. Vacancy, search, and prices in a housing market matching model. *J. Polit. Econ.* 98 (6), 1270–1292.
- Wheaton, W.C., Nechayev, G., 2008. The 1998–2005 Housing “Bubble” and the current “correction”: what’s different this time? *J. Real Estate Res.* 30 (1), 1–26.
- Yellen, J.L., 2005. Housing Bubbles and Monetary Policy. In: Presentation to the Fourth Annual Haas Gala, San Francisco, CA. <http://www.frbsf.org/our-district/press/presidents-speeches/yellen-speeches/2005/october/housing-bubbles-and-monetary-policy/051021.pdf>.

This page intentionally left blank

CHAPTER 12

Housing, Finance, and the Macroeconomy

Morris A. Davis*, **Stijn Van Nieuwerburgh[†]**

*Department of Finance and Economics, Rutgers Business School, Rutgers University, Newark, NJ, USA

[†]Department of Finance, Stern School of Business, New York University, NY, USA

Contents

12.1. Introduction	754
12.2. Stylized Facts	756
12.2.1 Levels of aggregates	756
12.2.2 Cross-sectional facts	761
12.2.3 Volatilities and correlations	764
12.3. Housing and the Business Cycle	767
12.4. Housing over the Life Cycle and in the Portfolio	773
12.4.1 A typical model	774
12.4.1.1 Utility	774
12.4.1.2 Choices	775
12.4.1.3 Constraints	775
12.4.1.4 Expectations for wages and prices	777
12.4.2 Calibration and estimation	778
12.4.3 Major findings in the portfolio literature	780
12.4.3.1 Housing and consumption patterns over the life cycle	780
12.4.3.2 The role of housing in the financial portfolio	783
12.4.3.3 Mortgage choice	785
12.5. Housing and Asset Pricing	787
12.5.1 Representative agent model	787
12.5.2 Risk sharing with housing collateral	790
12.6. The Housing Boom and Bust and the Great Recession	792
12.6.1 Exogenous house prices, no default	793
12.6.2 Exogenous house prices with default	794
12.6.3 Endogenous house prices, no default	796
12.6.4 Endogenous house prices with default	799
12.7. Housing Policy	800
12.7.1 Mortgage interest rate deductibility	801
12.7.2 Housing finance intermediaries	802
12.7.3 Housing and monetary policy	803
12.7.4 Mortgage modification programs	804
12.8. Conclusion	804
Acknowledgments	805
References	805

Abstract

In this chapter, we review and discuss the large body of research that has developed over the past 10-plus years that explores the interconnection of macroeconomics, finance, and housing. We focus on three major topics—housing and the business cycle, housing and portfolio choice, and housing and asset returns—and then review the recent literature that studies housing and the macroeconomy during the great housing boom and bust of 2000–2010. Our emphasis is on calibrated models that can be compared with data. In each section, we discuss the important questions, the typical set of tools used, and the insights that result from influential articles. Although great progress has been made in understanding the impact of housing outcomes on macroeconomic aggregates and vice versa, work remains. For example, economists recognize the importance of changing credit-market conditions in amplifying the volatility of house prices, but cannot explain the timing of these changes. At the end of the chapter, we discuss a new literature that assesses the macroeconomic effects and welfare implications of housing policies.

Keywords

Housing, Housing and macroeconomics, Housing and portfolio choice, Housing and asset pricing, Real estate boom and bust, Financial crisis, Housing crisis, Housing cycle, Real estate and business cycle

JEL Classification Codes

R00, R20, R31, D14, D31, D91, E21, E32, E44, E69

12.1. INTRODUCTION

Like disco and bell-bottomed pants, the study of housing has become fashionable among economists again. The tremendous boom (2000–2006) and bust (2006–2010) of housing markets and the subsequent financial crisis precipitated by an unforeseen surge in mortgage defaults have left many economists asking: What happened, why did it happen, and will it happen again? With this context in mind, the purpose of this chapter is to document what has been studied so far, what we think as economists we understand, and what we think we do not understand. Prior to the dramatic events of the past decade, economists investigated the interplay of housing and the macroeconomy for the obvious reasons that housing accounts for a large percentage of wealth and investment in housing accounts for a large fraction of overall economic activity. In addition, housing has some unique features that distinguish it from other assets. Specifically, (a) housing is infrequently traded and trades are subject to search frictions and large transaction costs, (b) the dividends that housing provide are unique to housing in the sense that only a structure can provide shelter and in the case of owner occupancy are hard to quantify, (c) the value of the asset class is enormous, and (d) the federal government interferes significantly in housing and mortgage markets. The sheer size of the housing and mortgage markets suggests that these peculiarities might affect macroeconomic outcomes and all other asset prices.

Many excellent articles have been produced on these topics, and it is impossible to cover all ground in one chapter. The articles and topics we write about almost all focus on the data and experiences in the United States and reflect our tastes in research. Our goal is to document the methods and explain the results of recent quantitative, mostly calibrated models produced by researchers in the fields of macroeconomics and finance. To this end, the chapter has eight sections. We start by highlighting key facts about housing in the United States. These facts typically serve as calibration or estimation targets; or, in some cases, they are the focus of specific research questions. Then, in order, we cover the topics of business cycles and housing; housing and portfolio choice in partial equilibrium models; and housing and asset prices. After this, we review a recent set of articles that aim to jointly explain business cycle facts, life-cycle portfolio facts, and equilibrium asset prices during the great housing boom and bust of the 2000–2010 decade. We distinguish between articles that fix house prices or take them as given and articles where house prices are determined endogenously inside the model. The penultimate section of the chapter describes the results of a small set of articles studying the impact and implications of US housing policy, with specific emphasis on the preferential tax treatment of owner-occupied housing and recent foreclosure relief efforts.

While we include quite a lot in this chapter, we do not cover everything. For example, we omit discussion of search frictions in housing markets¹; the relationship of housing and long-run demographic projections and the impact of housing on the long-run growth rates of output and consumption²; and articles with results that rely on agents having different expectations about the path of house prices.³ We also omit discussion of the active empirical literature that investigates the causes and consequences of the subprime mortgage crisis using techniques from applied microeconomics⁴ and articles investigating the role of mortgage finance in amplifying the housing boom and bust.⁵ Finally, this chapter is largely focused on the US experience. Home ownership rates and mortgage financing patterns differ across countries, and more research is needed to explain why these differences occur and how they may affect macroeconomic outcomes.⁶

¹ Several macroeconomists have contributed to this area. A small set of examples include [Albrecht et al. \(2007, 2010\)](#), [Head et al. \(2011\)](#), [Wong and Wright \(2011\)](#), [Head and Lloyd-Ellis \(2012\)](#), [He et al. \(2013\)](#), [Hedlund \(2014\)](#), [Landvoigt et al. \(2013a\)](#), [Ngai and Tenreyro \(2014\)](#), and [Piazzesi et al. \(2013\)](#).

² See, for example, [Mankiw and Weil \(1989\)](#) and [Davis et al. \(2014\)](#).

³ See, for example, [Piazzesi and Schneider \(2009\)](#) and [Burnside et al. \(2011\)](#).

⁴ Examples are [Mian and Sufi \(2009, 2011, 2012, 2014\)](#), [Mian et al. \(2010, 2013, 2014\)](#), and the references therein.

⁵ See, for example, [Keys et al. \(2009, 2010, 2012\)](#), [Piskorski et al. \(2010\)](#), and the review article by [Keys et al. \(2013\)](#).

⁶ For articles describing experiences of OECD countries, see [Catte et al. \(2004\)](#) and [Hirata et al. \(2013\)](#). [Aruoba et al. \(2014\)](#) note that source data on house prices for many countries are simply unavailable prior to 1990, hindering empirical analysis.

12.2. STYLIZED FACTS

12.2.1 Levels of aggregates

In this section, we highlight a set of stylized facts from US data that models of housing and the macroeconomy appropriate to explain outcomes in the United States should match. We begin by defining a set of first moments, or level variables, that are key to identifying parameters related to preferences and technology. Figure 12.1 shows the ratios of aggregate housing wealth to GDP and housing structures to GDP. Housing wealth is defined as the sum of housing structures and the market value of land, so the gap between the two series plotted in Figure 12.1 is equal to the ratio of the market value of land to GDP. These data are taken from Davis and Heathcote (2007), but similar data can be constructed from the Federal Reserve Board's Flow of Funds Accounts tables.⁷ According to these data, over the 1975–2013 period the average ratio of housing wealth to GDP is

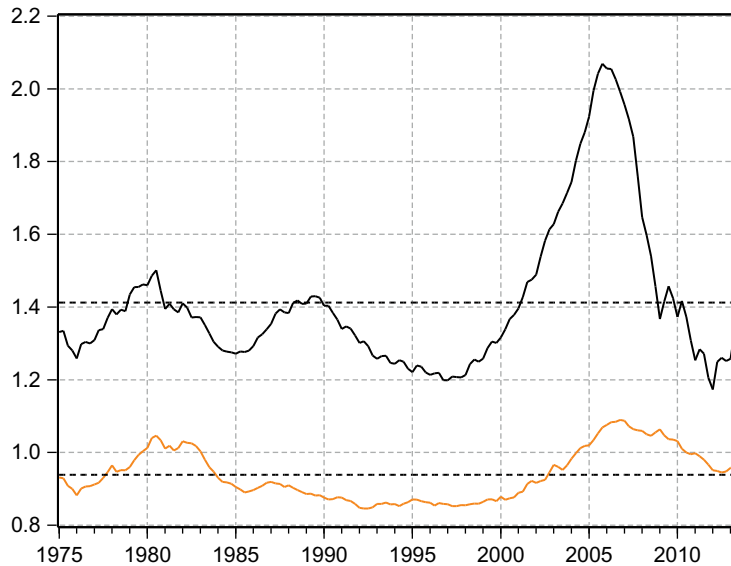


Figure 12.1 The figure plots the nominal value of housing (black), inclusive of land and structures, and the nominal value of structures (orange (light gray in the print version)), both relative to nominal GDP from the first quarter of 1975 to the third quarter of 2013. The housing and structures data are from Davis and Heathcote (2007) and are available for downloading at <http://www.lincolnst.edu/subcenters/land-values/price-and-quantity.asp> and the nominal GDP data are from the National Income and Product Accounts.

⁷ The data from Davis and Heathcote (2007) enforce that the capital gains from housing line up with changes in the Case–Shiller–Weiss price indices. This is not the case with the Flow of Funds data. The Davis and Heathcote (2007) data are available for downloading at <http://www.lincolnst.edu/subcenters/land-values/price-and-quantity.asp>.

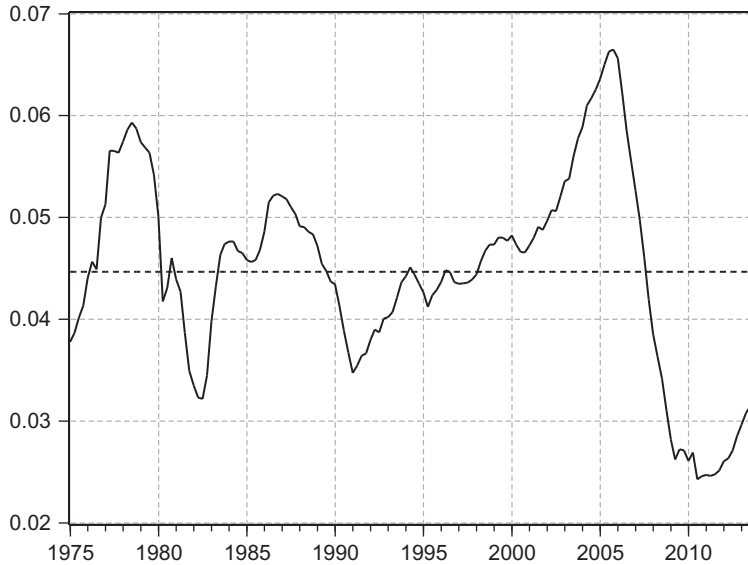


Figure 12.2 This figure plots the ratio of residential investment to GDP from the first quarter of 1975 to third quarter of 2013. These data are derived from Table 1.1.5 of the National Income and Product Accounts.

about 1.4 and the average ratio of the replacement cost of housing structures to GDP is 0.94. On average, the market value of land is roughly 45% of GDP, but the land share is volatile: At the height of the housing boom, the value of land was equal to 100% of GDP.

Figure 12.2 shows the ratio of residential investment to GDP over the same time period.⁸ The average value of this ratio is 4.5%. In most models, knowledge of the average value of the ratio of structures to GDP and residential investment to GDP is sufficient to pin down the implied rate of depreciation on residential structures. To see this, note that a capital accumulation equation for the real stock of structures implies

$$K_{t+1} = K_t(1 - \delta_K) + I_t.$$

If we assume a zero inflation rate and no trend to the relative price of structures, we can divide both sides by GDP at time t , Y_t , to uncover

$$\left(\frac{K_{t+1}}{Y_{t+1}}\right)\left(\frac{Y_{t+1}}{Y_t}\right) = \left(\frac{K_t}{Y_t}\right)(1 - \delta_K) + \frac{I_t}{Y_t}.$$

⁸ These data are derived from National Income and Product Accounts (NIPA) Table 1.1.5. Residential investment in the NIPA includes payments of brokers' commissions on the sale of homes. Although we do not do it here, authors occasionally remove these commissions from residential investment and adjust the estimate of the stock of structures accordingly.

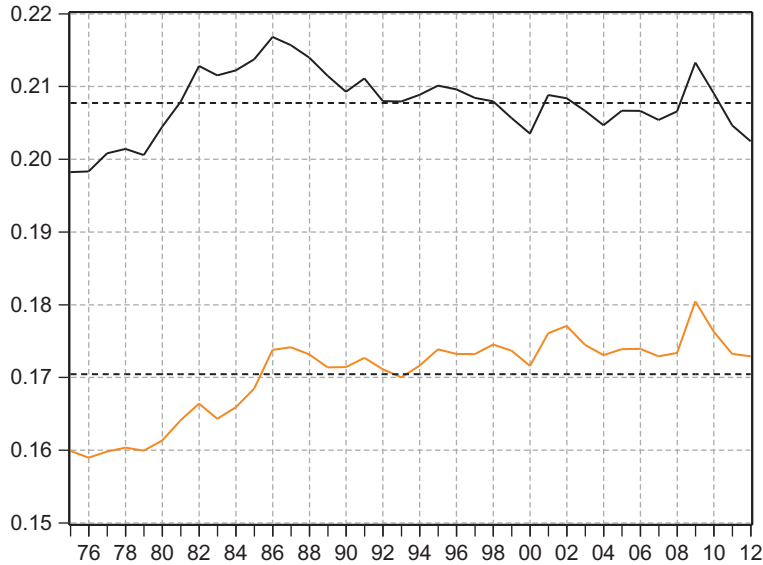


Figure 12.3 This figure plots the ratios of total spending on housing and utilities (black line) and housing (orange (light gray in the print version) line) to consumption expenditures on nondurable goods and services, from the first quarter of 1975 to third quarter of 2013. These data are derived from Table 2.4.5 of the National Income and Product Accounts.

Denote the steady-state ratios of structures to GDP, residential investment to GDP, and the growth rate of real GDP ($Y_{t+1}/Y_t - 1$) as K/Y , I/Y , and g_y , respectively. The model-implied rate of depreciation δ_K can be written as

$$g_y + \delta_K = \frac{I/Y}{K/Y}.$$

Taking I/Y as 0.045 and K/Y as 0.95, $g_y + \delta_K$ is 0.047. If we think the growth rate of real GDP inclusive of population growth is 3% per year, then this gives an estimate for δ_K of 1.7%. Most studies use a number in the vicinity of this estimate.

Some other first moments help to pin down parameters related to preferences. Figure 12.3 shows the aggregate ratio of expenditures on housing rents and utilities (black line) and housing rents (orange (light gray in the print version) line) to total consumption expenditures on nondurable goods and services.⁹ About 21% of aggregate consumption is accounted for by housing and utilities, 4% by utilities alone, and this percentage has remained constant despite real increases in consumption and rental prices over time. One typical objection to this evidence notes that, on average after 1975, more than 60% of measured aggregate spending on housing rents and utilities is imputed as rents

⁹ These data are derived from NIPA Table 2.4.5.

accruing to homeowners. Davis and Ortalo-Magné (2011) use microdata on renting households from the 1980, 1990, and 2000 Decennial Censuses of Housing to show that, across metropolitan areas and over time, the median ratio for renters of rental expenditures and utilities to household income is nearly constant at 24%. Although these results are not without controversy, researchers use the combination of evidence from Figure 12.3 and from Davis and Ortalo-Magné (2011) to justify preferences for housing and consumption that deliver constant expenditure shares on housing rents in the absence of borrowing constraints and frictions.

Finally, many researchers use data on price–rent ratios to help calibrate discount factors and to understand expectations about the future growth rate of rents and prices. An example of an estimate of aggregate price–rent ratios taken from Davis et al. (2008) is shown in Figure 12.4.¹⁰ This figure shows a slightly increasing ratio of prices to rents prior to 2000, a massive surge (2000–2006) and collapse in the ratio (2006–2010) during the housing boom and bust, and a return to trend after 2010. Of course, rents must be imputed to homeowners and therefore the reported level can vary from study to study depending on the imputation procedure. Therefore, Figure 12.5 plots five additional

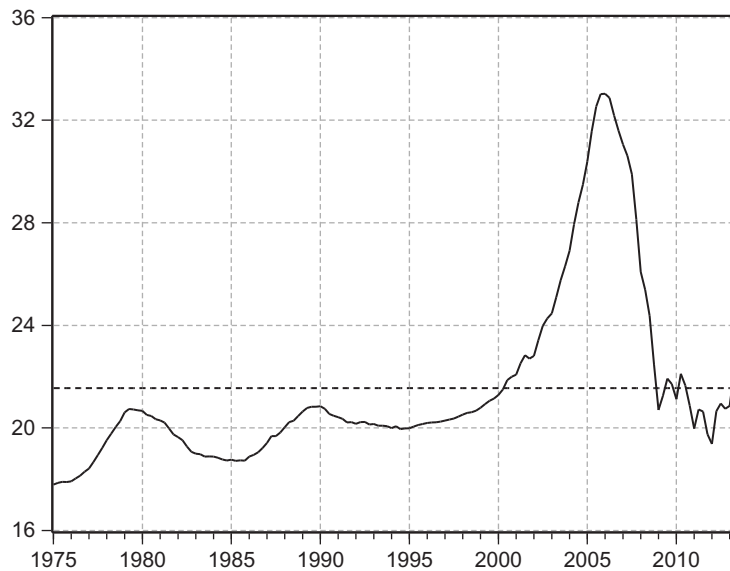


Figure 12.4 This figure plots the price–rent ratio for homeowners as derived by Davis et al. (2008) from the first quarter of 1975 to third quarter of 2013. These data are available for downloading at <http://www.lincolnst.edu/subcenters/land-values/rent-price-ratio.asp>.

¹⁰ These data are available at <http://www.lincolnst.edu/subcenters/land-values/rent-price-ratio.asp>.

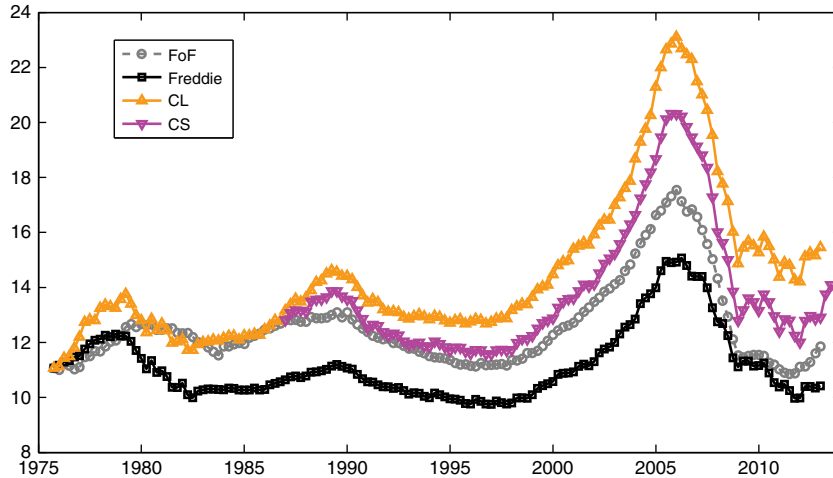


Figure 12.5 The figure compares five estimates of price-rent ratios from the fourth quarter of 1975 to the fourth quarter of 2013. “Flow of Funds” is the ratio of residential real estate wealth of the household sector from the Flow of Funds to aggregate housing services consumption from the National Income and Product Accounts. “Freddie Mac” is the ratio of the Freddie Mac Conventional Mortgage Home Price Index for purchases to the Bureau of Labor Statistics’ price index of shelter (which measures rent of renters and imputed rent of owners). “CoreLogic” is the ratio of the CoreLogic national house price index to the Bureau of Labor Statistics’s price index of shelter. “Case–Shiller” is the ratio of the Case–Shiller national house price index to the Bureau of Labor Statistics’s price index of shelter. The data are quarterly from the fourth quarter of 1975 or whenever first available until the fourth quarter of 2013.

estimates of the national price-rent ratio. The first one (circle markers) uses the ratio of aggregate housing wealth from the Flow of Funds divided by aggregate housing consumption from NIPA, and the next three series combine three different repeat-sales house price indices (the Freddie Mac index, squares; the CoreLogic index, up triangles; and the national Case–Shiller House Price Index, down triangles) with the shelter component of the Consumer Price Index of the Bureau of Labor Statistics.¹¹ All series display a similar pattern. The Flow of Funds (“FoF” in Figure 12.5) series and the Freddie Mac series (“Freddie” in Figure 12.5) display the least overall volatility, while the Case–Shiller series (“CS” in Figure 12.5) and the CoreLogic series (“CL” in Figure 12.5) display the largest boom and bust.

¹¹ Since both price and rental series are indices, we set the first observation (fourth quarter of 1975) of the price-rent ratio for each of these three series equal to the corresponding quarter’s observation for the Flow of Funds series. Since the Case–Shiller series only starts in the first quarter of 1987, we initialize this series at the first quarter of 1987 value for the Flow of Funds series. For an in-depth discussion of the properties of house price indices and repeat-sales methodology, see Ghysels et al. (2013).

12.2.2 Cross-sectional facts

Understanding differences across households with respect to choices and outcomes is often a goal of researchers in macroeconomics and finance. In this section, we discuss differences and disparities across households as they pertain to housing-related variables. Perhaps the most important dimension of heterogeneity in models of housing involves renting and owning. [Figure 12.6](#) shows the path of the home ownership rate since 1975. In lockstep with the changes to house prices over the 2000–2010 period, the home ownership rate displays a pronounced boom and bust: a four percentage point increase and a four percentage point decline. Each percentage point represents the experiences of approximately one million households.¹²

A second source of heterogeneity involves saving and lending: some households borrow to finance a home purchase and other households, some abroad, lend those funds. The aggregate quantity of US mortgage debt has increased significantly over time. The value of mortgages was equal to 20% of housing wealth in the 1950s, 30% by the mid-1970s, and 40% by the mid-1990s. Aggregate “loan to value” ratios remained constant during the housing boom. After house prices crashed, mortgage debt relative to housing wealth achieved its peak value of 62% in 2009. Household deleveraging (including defaults) alongside recovering property values has pushed the mortgage debt to housing wealth ratio back down to 50%. [Figure 12.7](#) also shows the ratio of mortgage debt to

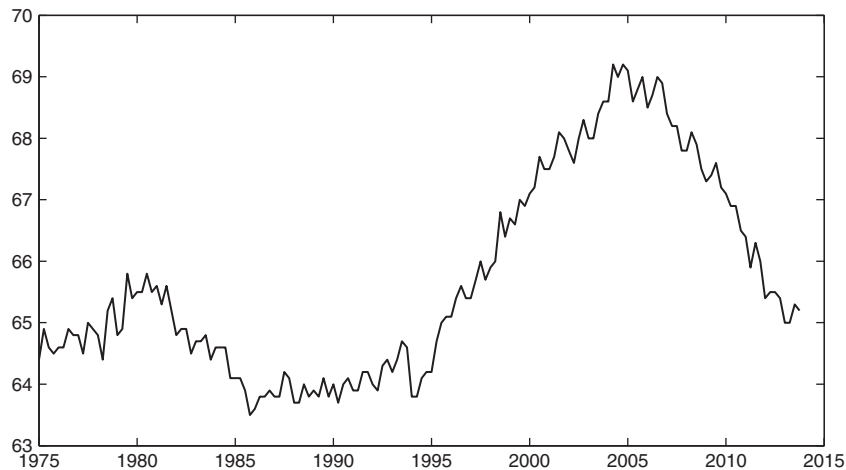


Figure 12.6 The figure plots the quarterly home ownership rate from the first quarter of 1975 to the fourth quarter of 2013. The data are from the US Department of Commerce: Census Bureau (FRED series ID RHORUSQ156N).

¹² As an aside, the home ownership rate increased dramatically between 1940, when it was in the low 40% range, and 1960, when it was in the low 60% range (see [Garriga et al., 2014](#)).

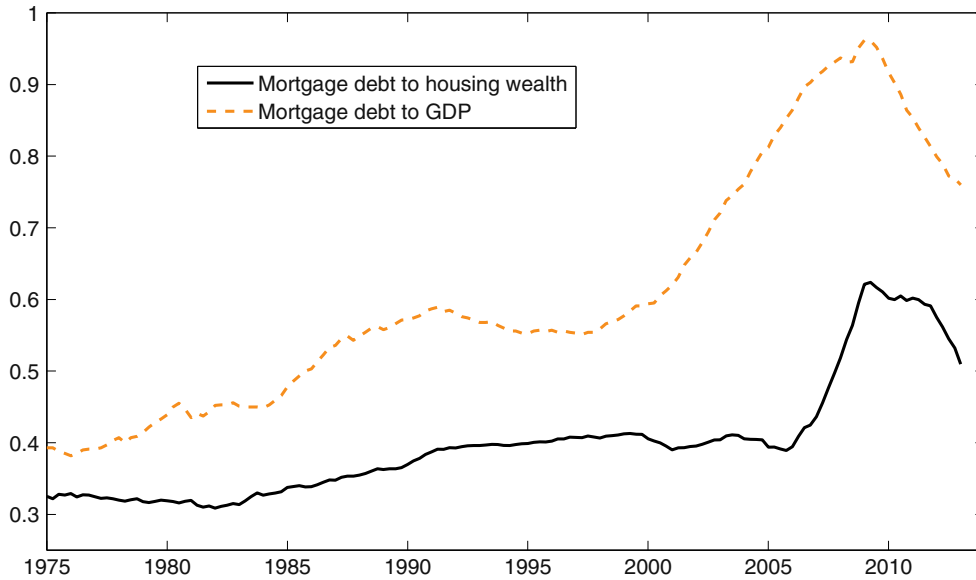


Figure 12.7 The figure plots mortgage debt relative to real estate wealth of the household sector (solid line) and mortgage debt of the household sector relative to GDP (dashed line) in the United States from the first quarter of 1975 to the first quarter of 2013. The data are from the Federal Reserve Board's Flow of Funds Accounts Tables B100.d and B103.d. Household real estate wealth excludes the real estate wealth of nonprofits, but includes the value of rental housing owned by the household sector (listed in Table 103.d but included in private business wealth in Table B100.d). Similarly, household mortgage debt includes the mortgage debt of the nonfinancial noncorporate sector. The GDP data are from the Bureau of Economic Analysis National Income and Products Accounts.

GDP. It tracks the ratio of mortgage debt to household wealth during the housing boom but shows a stronger deleveraging effect during the bust. Over the last 4 years, mortgage debt has fallen from 96% to 76% of GDP.

Next, we document a few important stylized facts about household portfolios using data from the Survey of Consumer Finances (SCF). [Figure 12.8](#) reports the home ownership rate by age for three different waves of the SCF: 2003, 2007, and 2010. In each SCF wave, home ownership rates rise with age until age 50 years and then stay flat until age 80 years. [Figure 12.8](#) also shows that home ownership rates fell at almost every age between 2007 and 2010, indicating that changes in the aggregate home ownership rate between 2007 and 2010 reflect changes across much of the population.

[Figure 12.9](#) shows the average household's net worth, defined as household assets less household liabilities, by age for homeowners and for renters. All data in this figure are in constant 2010 dollars. Clearly, average wealth is much higher for homeowners than for renters. In addition, homeowner wealth profiles have a clear "hump" shape over the life cycle, while wealth for renters is low and relatively constant with age. Combined with

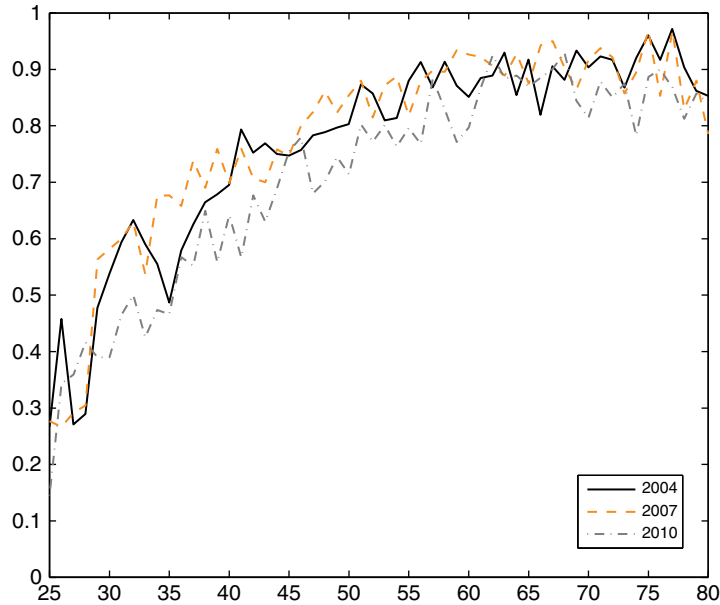


Figure 12.8 The figure plots the home ownership rate by age for the 2003, 2007, and 2010 waves of the Survey of Consumer Finances.

the previous graph, this figure suggests that homeowners decumulate financial wealth but not housing wealth late in life. This graph also suggests that homeowners and renters have quite different life experiences.

Figure 12.10 plots portfolio shares for major classes of assets by age using data from the 2010 SCF. The left panels show data for homeowners and the right panels show the data for renters.¹³ The top panels look at the share of housing, stocks, bonds, and retirement assets in total assets.¹⁴ The bottom panels plot home equity, stocks, bonds net of unsecured debt, and retirement assets as a share of net worth.¹⁵ The top-left panel shows that housing wealth accounts for most (90%) of the assets of homeowners early in life. Because most young homeowners take out a substantial amount of mortgage debt, home equity accounts for a somewhat smaller but still very large fraction of net worth (bottom left).

¹³ Homeowners are identified as households having positive housing wealth.

¹⁴ Retirement assets are difficult to split into stocks and bonds and so we leave them as a separate category. We include bank accounts with bonds and mutual fund holdings with stocks. The four shares sum to 1; for the calculation of these shares, we disregard the asset categories of vehicles, “other” financial assets, and business wealth.

¹⁵ Bonds are defined as the sum of bonds and bank accounts less credit card debt plus other financial assets less other unsecured debt. Home equity is defined as the sum of the values of “primary housing” and “other housing” less all mortgage debt. Stocks and retirement accounts are defined as before. By construction, the graphed shares sum to 1.

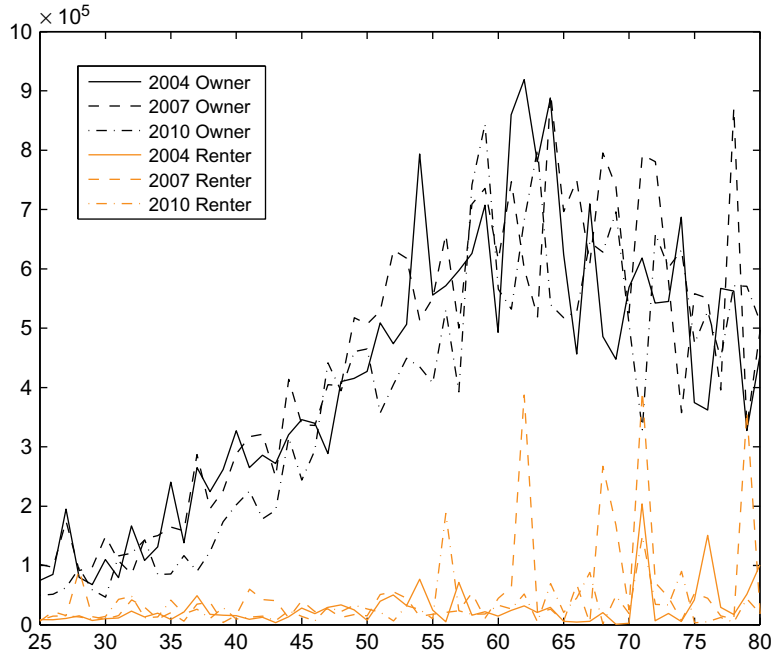


Figure 12.9 This figure plots average net worth by age for the 2003, 2007 and 2010 Survey of Consumer Finances. All nominal wealth estimates are deflated by the Consumer Price Index (base year 2010) published by the Bureau of Labor Statistics. Household wealth is expressed on a *per capita* basis by dividing by the number of adult members of the household. Wealth is computed as the sum of all assets minus the sum of all liabilities. Assets are pension assets, bank accounts, IRAs, bond holdings, mutual fund holdings, stock holdings, the value of the primary residence if owned, other housing wealth, business wealth, other financial wealth, and vehicles. Liabilities are credit card debt, mortgage debt for the primary residence, debt for other property, and other debt.

As homeowners age, housing falls to about 50% of total assets by age 60 years, and financial wealth becomes a larger share of assets and net worth. Late in life, retirement assets are depleted, and stocks, bonds, and housing all account for a significant fraction of wealth. In contrast, renters have little stock market wealth; most of their wealth is in retirement assets and in bonds.

12.2.3 Volatilities and correlations

Macroeconomic researchers studying the properties of business cycle models typically force the models to match key first moments, such as the first moments we discussed earlier in the chapter, and then informally evaluate model performance by judging how well the model can replicate important second moments. This procedure is in the spirit of [Kydland and Prescott \(1982\)](#), who ask if a macroeconomic model can simultaneously be consistent with the long-run growth facts of [Kaldor \(1957\)](#) and match the business cycle facts of

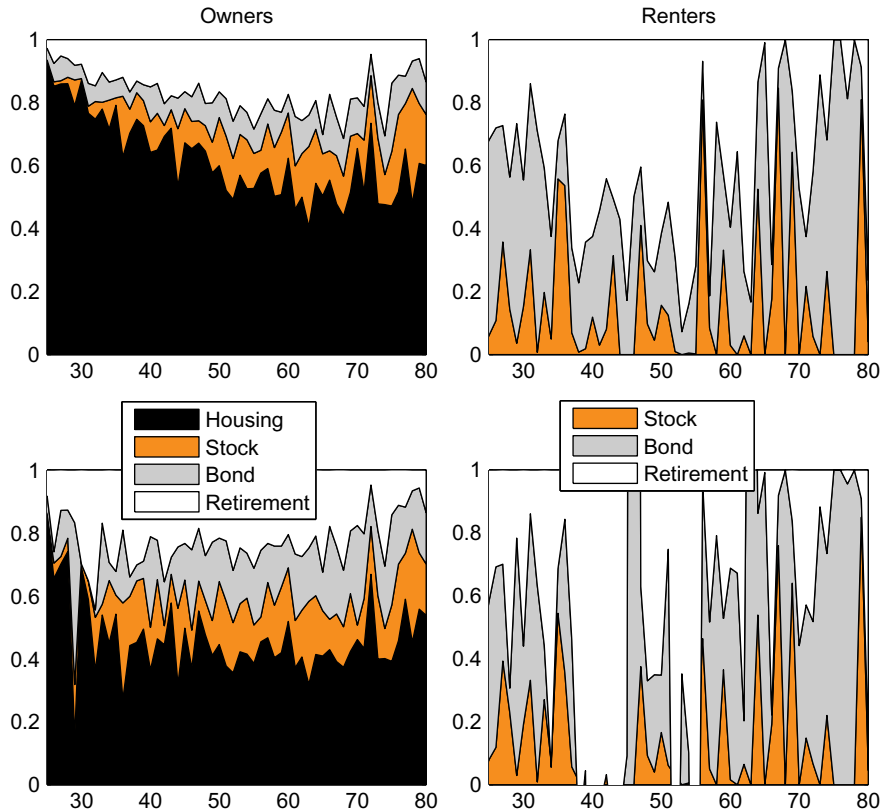


Figure 12.10 The figure plots portfolio shares by age for homeowners (left panels) and renters (right panels) for the 2010 Survey of Consumer Finances wave. The top panel plots the share of four asset categories: housing, stocks, bonds, and retirement accounts. Retirement assets are difficult to split into stocks and bonds, and so we leave them as a separate category. We include bank accounts as bonds and mutual fund holdings with stocks. The four shares sum to 1. That is, for the calculation of these shares, we disregard the remaining asset categories of vehicles, other financial assets, and business wealth. The bottom panels plot home equity, stocks, bonds net of unsecured debt, and retirement assets as a share of net worth. In these panels, bonds are defined as the sum of bonds and bank accounts less credit card debt plus other financial assets less other unsecured debt. Home equity is defined as the sum of the value of all housing owned less the amount of all mortgage debt owed. Stocks and retirement accounts are defined as in the top panels, and the four shares sum to 1.

Burns and Mitchell (1946).¹⁶ Through the appropriate choice of functional forms for production and utility functions, many macro models will, by definition, be consistent with first moments; this implies model evaluation should focus on second moments. [Table 12.1](#)

¹⁶ For example, the ratios of consumption, investment, and capital to output are stable, on average over many years, but are all positively correlated over the business cycle.

Table 12.1 Properties of selected detrended US macroeconomic data, first quarter of 1955 to third quarter of 2013

Variable X		Standard deviation	Relative standard deviation	Correlation of variable X_s and GDP_t						
				$s = t - 3$	$t - 2$	$t - 1$	t	$t + 1$	$t + 2$	$t + 3$
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
(a)	GDP	1.54	1.00	0.40	0.64	0.86	1.00	0.86	0.65	0.42
(b)	Consumption	0.85	0.55	0.49	0.67	0.81	0.84	0.75	0.59	0.41
(c)	Nonresidential investment	4.74	3.07	0.13	0.36	0.61	0.81	0.87	0.82	0.70
(d)	Residential investment	9.98	6.47	0.67	0.75	0.76	0.66	0.45	0.21	−0.02
(e)	House prices ^a	4.16	2.70	0.47	0.53	0.55	0.52	0.46	0.41	0.35
(f)	Durables quantities	4.49	2.91	0.51	0.67	0.79	0.82	0.65	0.44	0.21
(g)	Durables prices	0.93	0.60	0.13	0.05	−0.04	−0.15	−0.24	−0.30	−0.35

^aData begin in the first quarter of 1975.

Data are quarterly. All data except the house price data are from the National Income and Product Accounts (NIPA) as produced by the Bureau of Economic Analysis. The house price data combine data from the Federal Home Finance Agency House Price Index (1975–1986) and the Case–Shiller–Weiss index as made available by Macro-markets (1987–2013). All variables have been logged and Hodrick–Prescott-filtered with smoothing parameter $\lambda = 1600$. Real house and durable prices are computed as the nominal price index divided by the price index for consumption of nondurable goods and services.

shows some of these second moments for US data—standard deviations and correlations—that researchers try to match. Generating these moments requires, as a first step, removing trends in the data. The typical procedure to remove those trends has been to use the Hodrick–Prescott filter.¹⁷

Inspection of Table 12.1 yields several important stylized facts. First, consumption, nonresidential investment, residential investment, and spending on durable consumption goods are all positively contemporaneously correlated with GDP (column 6). The positive correlation of these major macroeconomic variables is considered a key property of business cycles. Second, consumption is about half as volatile as GDP (cell b2 of the table); nonresidential investment is three times more volatile than GDP (c2); residential investment is more than twice as volatile as nonresidential investment (d2); and house prices are more than 2.5 times as volatile as GDP (e2).¹⁸ Finally, the highest correlation of nonresidential investment and GDP occurs when GDP is lagged once relative to nonresidential investment (c7); and the highest correlation of residential investment and GDP occurs when residential investment is lagged by one or two quarters (d4 and d5).¹⁹ Thus, residential investment leads business investment by about two quarters.

12.3. HOUSING AND THE BUSINESS CYCLE

The cyclical nature of housing has been a topic of interest for decades, and many economists have written on the topic. An active empirical literature studies the lead–lag relationship of housing with other macroeconomic aggregates (e.g., Green, 1997; Leamer, 2007; Ghent and Owyang, 2010), and the relationship of house prices and housing wealth to consumption (e.g., Muellbauer and Murphy, 1997; Davis and Palumbo, 2001; Case et al., 2005).²⁰ In this section, we focus on one specific branch of the literature: equilibrium aggregate models in the spirit of Kydland and Prescott (1982)—that is, “real business cycle” (RBC) models, where housing prices and quantities are endogenously determined inside the model alongside aggregate consumption, investment, and output, and aggregate market clearing conditions are defined as part of the equilibrium.²¹ We focus on these models because they reflect our tastes and research experience, but also because many macroeconomists currently studying housing use a similar framework.

¹⁷ The Hodrick–Prescott filter removes a stochastic trend (see Hodrick and Prescott, 1997 for details).

¹⁸ Almost identical results are obtained for every statistic when real house prices are replaced with the price–rent ratio (not shown).

¹⁹ Residential investment does not lead GDP in all countries (see Kydland et al., 2012).

²⁰ See Cooper and Dynan (2013) for a recent summary of that literature.

²¹ For example, we would not consider the models of Topel and Rosen (1988) and Grenadier (1995) to be part of this literature, since in both of those models many aggregate variables are determined outside the model and aggregate market clearing conditions for all variables are not specified.

The first class of RBC models where housing variables entered as objects of interest were models with home production (Benhabib et al., 1991; Greenwood and Hercowitz, 1991). Those models are two-sector extensions of the canonical RBC model of Kydland and Prescott (1982). The key extension of these models relative to the original RBC model is that households are assumed have three uses of time: market work and leisure, as in the standard model, and work at home. Households combine work at home with a stock of home capital, subject to a productivity shock to home output, to produce a good called “home consumption,” which is complementary in utility with leisure and with consumption purchased in the market. A justification for this approach (see McGrattan et al., 1997) is that in time-use surveys, households on average spend about 25% of discretionary time on activities that can be classified as home work. When authors in this literature calibrate their models, they set the stock of home capital equal to the sum of the stock of residential structures (housing less land) and the stock of durable goods and set gross investment in home capital equal to investment in residential structures and spending on consumer durables (see Greenwood et al., 1995; McGrattan et al., 1997).

In the home production literature, it is typically assumed that home capital and market capital have the same price per unit (except for adjustment costs), and that home capital can be modeled as a simple aggregate of durable goods and residential structures. Figure 12.11, which graphs real house prices, durable goods prices, and the price of non-residential fixed investment over the 1975–2013 period, shows that these assumptions are at odds with the data.²² The top panel of the figure plots the raw data, while the bottom panel shows the logged, Hodrick–Prescott–filtered data. The top panel shows that the real price of durable goods and business investment has been falling rapidly, while the real price of housing has been rising slowly. The bottom panel shows that cyclical durable goods prices and nonresidential investment prices are highly correlated, but the cyclical price of housing is more volatile than and is not correlated with either series.

These price data suggest that the productivity shocks affecting the production of housing may be quite different from shocks affecting production in other sectors of the economy.²³ To see this, consider the following simple two-sector economy. In the first sector (subscript c), a general good is produced that can be used for consumption or business investment. In the second sector (subscript h), residential structures used for housing are produced. Firms in each sector rent capital K and labor L from households to produce output Y according to the functions below:

²² All prices have been converted to real using the NIPA price index for consumption of nondurable goods and services.

²³ “Productivity shocks” allow the level of output to vary even when the quantity of inputs is held fixed. These shocks capture, in a reduced-form sense, changes to the methods by which firms manage and organize inputs (e.g., logistics, management structure) to more efficiently make output.

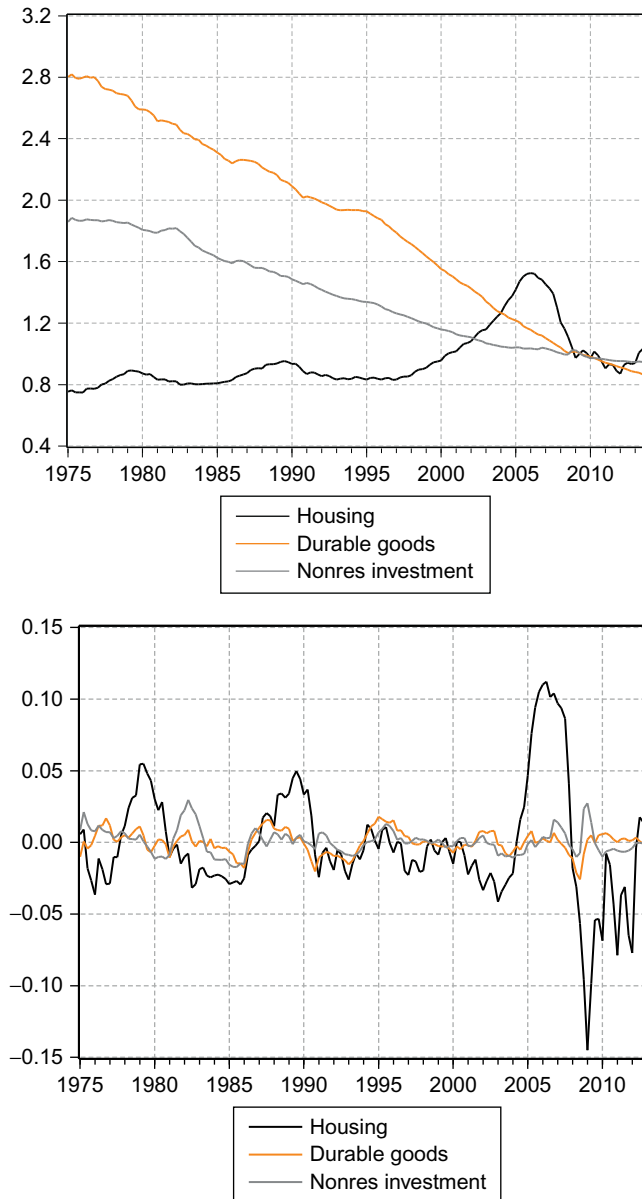


Figure 12.11 Relative prices (2009 prices correspond to 1.0), first quarter of 1975 to third quarter of 2013. The price index for durable goods and that for nonresidential investment are from the National Index and Product Accounts. The house-price data combine data from the Federal Home Finance Agency House Price Index (1975–1986) and the Case-Shiller-Weiss index as made available by Macromarkets (1987–2013). All prices have been converted to real prices by deflating them using the National Index and Product Accounts price index for consumption of nondurable goods and services. The top panel shows the raw data and the bottom panel shows the data after taking logarithms and applying the Hodrick–Prescott filter.

$$\begin{aligned} Y_c &= z_c K_c^\alpha L_c^{1-\alpha}, \\ Y_h &= z_h K_h^\alpha L_h^{1-\alpha}, \end{aligned}$$

where z_c and z_h are sector-specific productivity shocks. Normalize the price of consumption to 1 and denote the price of one unit of residential investment as p_h , the price of one unit of capital as r , and the price of one unit of labor as w . The profit maximization problems for the firms in each sector are

$$\begin{aligned} \max_{K_c, L_c} \quad & z_c K_c^\alpha L_c^{1-\alpha} - rK_c - wL_c, \\ \max_{K_h, L_h} \quad & p_h z_h K_h^\alpha L_h^{1-\alpha} - rK_h - wL_h, \end{aligned}$$

implying first-order conditions for optimal capital and labor utilization for firms in each sector of

$$\begin{aligned} r &= z_c \alpha \left(\frac{K_c}{L_c} \right)^{\alpha-1} = p_h z_h \alpha \left(\frac{K_h}{L_h} \right)^{\alpha-1}, \\ w &= z_c (1-\alpha) \left(\frac{K_c}{L_c} \right)^\alpha = p_h z_h (1-\alpha) \left(\frac{K_h}{L_h} \right)^\alpha. \end{aligned}$$

These two equations imply that we can solve for the relative price of residential structures in equilibrium as

$$p_h = \frac{z_c}{z_h}. \quad (12.1)$$

In this simple two-sector model, where we have assumed the capital share of production is the same in both sectors, the relative price of residential structures is equal to the ratio of the two productivity shocks.²⁴ This result is the reason why authors interested in the cyclical behavior of residential investment and house prices have modeled the production of housing in some detail inside a business-cycle model.

The first article to model shocks affecting the production of housing is by [Davis and Heathcote \(2005\)](#). The model of Davis and Heathcote is similar in spirit to the simple two-sector model described above. The key difference is that Davis and Heathcote design the production side of the model such that all the inputs and key parameters are identifiable using available data. Other articles that have adopted a similar specification for housing production are those of [Kahn \(2008\)](#), [Iacoviello and Neri \(2010\)](#), [Kiyotaki et al. \(2011\)](#), and [Dorofeenko et al. \(2014\)](#) to name just a few recent examples.

In [Davis and Heathcote \(2005\)](#), one set of firms produces “intermediate goods” (similarly to what is done in the article by [Hornstein and Praschnik, 1997](#)) and a second set of firms transforms these intermediate goods into final goods. There are three types of firms

²⁴ Adjustment costs can introduce a wedge between the price producers receive as computed in (12.1) and the price households pay for additional installed unit of capital (see [Fisher, 1997](#) for more details).

producing intermediate goods—a first type produces construction output (subscript b), a second type produces manufacturing output (subscript m), and a third type produces services output (subscript s). Call each of these types of firms “industries.” Output x from each industry is made from capital K and labor N rented from households and is subject to an industry-specific productivity shock z according to

$$x_{it} = z_{it} K_{it}^{\theta_i} N_{it}^{1-\theta_i} \quad \text{for } i = \{b, m, s\}.$$

Notice that the capital share θ_i is allowed to vary by industry. Davis and Heathcote (2005) identify θ_i for each industry and then use data on K_i and N_i from the Bureau of Economic Analysis to estimate the time-series values of z_{it} for each industry. These z_{it} are the only shocks in the Davis and Heathcote model.

Continuing, Davis and Heathcote assume three types of firms that produce final goods. The first type of firm produces a good (subscript c) that can be used by households for either consumption or business investment. The second type of firm produces residential investment (subscript d). The third type, discussed below, produces housing (subscript h). The first two types of firms that produce final goods use construction, manufacturing, and services goods as inputs. These firms produce their output (y) according to

$$y_{jt} = b_{jt}^{B_j} m_{jt}^{M_j} s_{jt}^{S_j} \quad \text{for } j = \{c, d\}.$$

B_j , M_j , and S_j are the shares of construction, manufacturing, and services value-add for residential investment ($j = d$) and consumption and business investment ($j = c$).

Davis and Heathcote identify the production-function parameters B_j , M_j , and S_j using data from the Input–Output tables of the NIPA. Since all three intermediate goods are used in both types of final goods, shocks to all three industries affect the production of both types of final goods. To the extent that residential investment is more construction intensive than consumption and business investment, $B_d > B_c$, the shocks affecting the construction sector will be relatively more important for determining residential investment. Thus, residential investment will have a price different from that of other goods depending on the properties of the z_{it} terms and the values of B_j , M_j , and S_j for $j = c, d$.

The third type of firm that produces final goods combines residential investment and new land to make housing. Denote x_d and x_l as residential investment and land that the housing-producing firms buy. The amount of new housing produced y_h is given by

$$y_{ht} = x_{lt}^\theta x_{dt}^{1-\theta}. \quad (12.2)$$

Davis and Heathcote assume that the amount of new land available in the economy is fixed at 1 each period. This assumption makes possible a closed-form solution for the total quantity of housing in the economy as a function of past investments in structures

Table 12.2 Business cycle properties of the Davis and Heathcote (2005) model

Variable		Data	Davis and Heathcote model
Standard deviations relative to GDP			
(a)	Hours worked	1.01	0.41
(b)	Nonresidential investment	2.30	3.21
(c)	Residential investment	5.04	6.12
(d)	House prices	1.37	0.40
Period t correlations			
(e)	Residential and nonresidential investment	0.25	0.15
(f)	Residential investment and house prices	0.34	− 0.20

All results and data in this table are taken from Table 10 in Davis and Heathcote (2005). The standard deviations and correlations in this table do not exactly match those in Table 12.1. Davis and Heathcote (2005) use annual data over a different sample range (1948–2001); they filter the data using the Hodrick–Prescott filter with smoothing parameter $\lambda = 100$; and they use different source data for house prices.

after accounting for depreciation and land accumulation. The inability to produce new land acts like an adjustment cost to the production of new housing.²⁵

For the household side of the model, the Davis and Heathcote model assumes a representative agent who receives utility from market consumption c_t , leisure $1 - L_t$, and housing H_t of the form

$$\frac{(c_t^{\mu_c} H_t^{\mu_h} (1 - L_t)^{1 - \mu_c - \mu_h})^{1 - \sigma}}{1 - \sigma}.$$

In this expression, total hours worked L_t is the sum of the hours worked in the construction, manufacturing, and services sectors.

Table 12.2 compares some key moments from the data and from simulations of the Davis and Heathcote model. Row (a) shows that the model underpredicts the volatility of hours worked. This finding is not surprising. Despite the fact that Davis and Heathcote model three productivity shocks, all activity in the model occurs in the market, so the three productivity shocks once aggregated should have, roughly speaking, the same properties as the one productivity shock of the canonical RBC model, and that model systematically underpredicts the volatility of hours worked. Rows (b) and (c) show the relative standard deviations of nonresidential and residential investment. Relative to the classic results from the home-production literature such as Gomme et al. (2001), the model has great success replicating the fact that residential investment is about twice as volatile as nonresidential investment. Additionally, the model replicates the positive contemporaneous correlation of nonresidential and residential investment (row (e)). This result arises from the fact that land acts as an adjustment cost to building new housing

²⁵ While the assumption of one unit of land available for development in each period is convenient, it is not clear what this quantity should be. Ultimately, the quantity of new land available for development determines the aggregate elasticity of supply of housing. More research is needed on this topic.

rapidly. As noted by Fisher (1997), these kinds of adjustment costs are necessary to generate positive co-movement of residential and nonresidential investment.

The model fails to match the housing data along three dimensions. First, residential investment tends to lead GDP and nonresidential investment tends to lag GDP. The model fails to replicate this finding. Second, the model underpredicts the volatility of house prices (row (d)).²⁶ Third, the model predicts a negative correlation of residential investment and house prices, whereas in the data the correlation is positive (row (f)).

The fact that the model predicts that house prices and residential investment are negatively correlated is likely due to the nature of shocks hitting the model.²⁷ Consider again the result of the simple two-sector model without adjustment costs, $p_h = z_c/z_h$. When residential investment has relatively high productivity and z_h is relatively high, house prices fall. But when z_h is high, it is a productive time to build housing. A simpler way to say this is that people should build housing when it is cheap to do so. Ultimately, there are two offsetting effects in the model of Davis and Heathcote making the analysis more complicated. First, land acts as an adjustment cost, and the fixity of new land available for new housing boosts house prices when building activity is high. Second, income effects can offset price effects. Referring again to the simple model, we find when z_c is high and house prices are relatively high, income is also high (since z_c is used to produce consumption and business investment, most of GDP). When income is high, households want more of everything, including housing.

Some progress has been made on fixing the three counterfactual findings of Davis and Heathcote (2005), but more work needs to be done. Fisher (2007) shows that when housing is included as a separate capital stock in the market production function, the lead-lag properties of the model are improved. Dorofeenko et al. (2014) show that the inclusion of “risk shocks” to the production of housing can help boost the volatility of house prices, but at the expense of other model statistics. Kydland et al. (2012) demonstrate that properties of the pricing of fixed-rate mortgages (FRMs) may be critical to understanding why residential investment leads GDP. Favilukis et al. (2011), whose work is discussed in more detail later in this chapter, generate considerably more volatile house prices in a heterogeneous-agent equilibrium model with collateral constraints. It also improves on the lead-lag relationship between residential investment and GDP and generates a positive correlation between residential investment and house prices.

12.4. HOUSING OVER THE LIFE CYCLE AND IN THE PORTFOLIO

A substantial literature researches the role of housing in households’ portfolios of assets. On average over 1952–2013, housing wealth accounted for 35% of household assets and

²⁶ The relative volatility of house prices is much lower in Table 12.2 than in Table 12.1 owing to differences in sample periods and source data. See the notes to Table 12.2 for details.

²⁷ The model of Fisher (1997) also predicts a negative correlation of house prices and residential investment.

40% of household net worth (assets minus liabilities), while home equity (housing wealth minus mortgage debt) accounted for 23% of assets and 26% of net worth.²⁸ As shown in Figures 12.6, 12.8, and 12.10, two-thirds of households in the United States own their home, and for most home-owning households, housing accounts for a substantial portion of total wealth.

Housing is not only an important asset in the portfolio, it also has several features that make it different from investments in financial assets. First, it is illiquid in the sense that changing the quantity of housing may take time and/or requires incurring substantial transaction costs. Second, it is indivisible: a limited assortment of types and sizes are available for purchase at any time (including a minimum size). Third, home ownership and housing consumption are typically intimately linked. Most households own only one home and live in the house they own. Fourth, housing represents the main source of pledgeable capital against which households can borrow. Investment in housing is much more leveraged than investments in other financial assets, and the value of owned housing limits the amount of leverage in households' portfolios. Fifth, housing is tied to a particular labor market: people usually live near where they work.

For the past 15 years or so, researchers have developed relatively simple life-cycle decision models with rational, forward-looking, optimizing agents able to reproduce systematic differences in the data on the home ownership, housing, and portfolio choices of various groups of households—sorted by tenant status (owner vs. renter), age, income, and net worth. With this in mind, we proceed by describing a model environment representative of many of the models studied in this literature and describe the typical calibration of key parameters and processes. Finally, we discuss in some detail the setup and results of many key articles in this now mature literature.

12.4.1 A typical model

12.4.1.1 Utility

Most authors assume that finite-lived households receive utility from nonhousing consumption c , the quantity of housing services h , and whether or not those services are acquired through owning $o = 1$ or renting $o = 0$, such that per-period utility can be written as $u(c, h, o)$. In each period, households act to maximize the present discounted value of remaining expected utility subject to constraints we discuss later. That is, if the household dies with certainty in T periods, in the current period $t = 0$ households maximize

²⁸ The data are from the Flow of Funds for the first quarter of 1952 until the first quarter of 2013, Table B100. The calculations reassign the portion of private business wealth in the household table that reflects households' ownership of real estate available for rent. The same is done for the corresponding mortgages. This reclassification is done on the basis of information in Table B103.

$$\sum_{t=0}^T \beta^t \pi(t) E_0[u(c, h, o)]. \quad (12.3)$$

In the equation above, β is the rate at which households discount future utility, $0 < \pi(t) < 1$ is the age-dependent probability of survival to period t , and E_0 denotes an expectation taken in the current period. Occasionally, authors assume that households have a bequest motive in which they or their offspring receive utility if they die with positive wealth. In that case, the net present discounted value of utility has one additional term equal to some payoff from wealth remaining at death.

The functional form for per-period utility can be generalized as a constant relative risk aversion utility over a joint consumption bundle with risk aversion parameter $\sigma \geq 1$, and where the consumption bundle is a constant elasticity of substitution (CES) aggregate over housing and nonhousing consumption, with intratemporal elasticity parameter $-\infty < \rho < 1$ and budget share parameter α :

$$u(c, h, o) = \frac{(\alpha c^\rho + (1 - \alpha) \xi_o h^\rho)^{\frac{1-\sigma}{\rho}}}{1 - \sigma}. \quad (12.4)$$

The parameter ξ captures the extra utility from housing services that are owned rather than rented: $\xi_1 > \xi_0$.

12.4.1.2 Choices

In each period of the model, households make a set of continuous choices and one discrete choice. The continuous choices are related to consumption, the quantity of financial assets, and the quantity of housing services. Studies differ in the treatment of financial assets. Some authors lump together all financial assets into a net position; others allow for one positive financial asset and one negative financial asset, typically a mortgage; others allow for a set of positive financial assets such as stocks and bonds; and finally some studies allow households to choose among multiple types of mortgages—for example, FRMs or adjustable-rate mortgages (ARMs). The discrete choice is whether or not to receive housing services by owning or renting.

12.4.1.3 Constraints

Each period, households are subject to a basic budget constraint, a net worth or home equity constraint, and one or more constraints about housing choices. Below, we consider the case of renters remaining renters and homeowners remaining homeowners. When homeowners in $t - 1$ become renters in t and renters in $t - 1$ become homeowners in t , constraints change in the expected ways.

Denote labor income as w_t , financial assets (consolidated for convenience) as a_t , the rate of return on financial assets as r_t , and the rental price per unit of housing as q_t . In all

studies income varies over time. In most studies, house prices vary over time. The basic budget and net worth constraints for households that were renters in the previous period $o_{t-1} = 0$ and choose to rent in the current period have the intuitive forms

$$c_t + q_t h_t + \tau_t + a_t = w_t + a_{t-1}(1 + r_{t-1}), \quad (12.5)$$

$$a_t \geq \underline{a}_t, \quad (12.6)$$

$$h_t \leq \bar{h}. \quad (12.7)$$

The first equation is simple accounting—out of available resources, whatever is not spent on consumption, taxes τ_t , or rental housing is saved. For renters, taxes paid is usually a simple function of age, income, and assets. The second equation imposes some discipline on spending. Consider what would happen in the period before death, $T - 1$, without a constraint like this: each household would spend as much as possible on consumption and housing. In many articles, $\underline{a}_t$ is set to 0 in each period, and in a few articles, $\underline{a}_t$ is determined endogenously such that the household can repay any debt with probability 1. The final constraint is that rental housing cannot be too large.

The budget and net worth constraints for households that owned their home in the previous period and choose to own their home in the current period are more complicated. The simpler models tend to have the form

$$c_t + p_t h_t + \tau_t + mc + a_t = w_t + a_{t-1}(1 + r_{t-1}) + p_t(1 - \delta)h_{t-1}, \quad (12.8)$$

$$mc = \zeta p_t h_{t-1} \text{ if } h_t \neq h_{t-1}, 0 \text{ otherwise}, \quad (12.9)$$

$$h_t \geq \underline{h}, \quad (12.10)$$

$$a_t \geq -(1 - \phi)p_t h_t. \quad (12.11)$$

The first equation is accounting. It states that out of available resources (income, financial assets, and housing wealth after accounting for depreciation δh), whatever is not spent on consumption, taxes, or housing is saved. The amount spent on housing is the value of the house bought plus moving costs mc . As shown in Equation (12.9), moving costs, typically assumed to be a fixed percentage ζ of the value of the house being sold, are incurred if the size of the owned house changes between periods—that is, $h_t \neq h_{t-1}$. Some authors include a fixed component in the moving cost. The moving costs allow models to capture the fact that households move infrequently.

The term $r_{t-1}a_{t-1}$ reflects net interest received or paid (if negative) by households. The embedded assumption is that homeowners can borrow and lend at the same rate r , so a is the value of all savings net of any debt owed. In this stylized budget constraint, there are no costs to adjusting the quantity of debt or assets, so households can vary their mortgage or other financial assets costlessly over time as long as their debt level is not too high. Some authors relax this assumption and allow for participation costs or other costs when households change the size of their mortgage (refinance) or adjust their portfolio

between stocks and bonds. A sizeable literature studies costs households pay to participate in stock markets (e.g., [Vissing-Jorgensen, 2002](#)).

Equation (12.10) specifies that owned housing cannot be too small. Equation (12.11) is a key constraint which governs maximum homeowner debt. Homeowners can borrow up to $1 - \phi$ fraction of their house value; ϕ is sometimes called the down-payment fraction. The down-payment constraint (or, transformed, the loan-to-value constraint) and the minimum-sized owned housing constraint allow models to match low home ownership rates for the young: the young need to save to afford the down payment on their first house. Models of this class usually rule out default.²⁹

The budget constraints for homeowners, taxes paid, and laws of motion for wealth and debt can be more complicated in studies of how different mortgage types affect the demand for housing. Some authors assume that homeowners cannot borrow and save at the same rate, and cannot add to their mortgage debt without incurring a cost. In these articles, the constraint regarding the maximum quantity of mortgage debt is typically considered only for first-time home buyers, or for homeowners moving to a different house. That is, homeowners are (usually) not forced to move if the value of their house unexpectedly declines and their loan-to-value ratio increases. The simple model also assumes that mortgage debt has one-period maturity, resembling ARMs. Some authors explicitly allow for long-term mortgages, usually modeled as perpetuities whose coupon payments are set to match the average duration of FRMs. Authors who study mortgage design also separately model interest and principal repayment schedules, and mortgage interest on debt is assumed to be tax deductible. Finally, some authors add debt-to-income constraints in addition to loan-to-value constraints.

12.4.1.4 Expectations for wages and prices

Earnings in these models are almost always stochastic around a known age-specific hump-shaped profile. Households are assumed to know the process for wages, but not the future realizations of the shocks. Assumptions about house prices vary. Sometimes house prices are fixed or grow at a predetermined rate, but most of the time they are stochastic. If they are random, typically the level is assumed to be persistent, either a random walk (possibly with time-varying drift as in [Corradin et al., 2014](#)), or a first-order autoregressive process with a value for the autoregressive coefficient near 1. In models where house prices are not determined endogenously, the process for calibrating the covariance of shocks to income and shocks to house prices varies across authors. Households in these models are forward-looking with rational expectations.

²⁹ We discuss models that allow for default later in the chapter. In these models, the down-payment constraint limits the likelihood of default, but does not rule it out.

12.4.2 Calibration and estimation

These models are estimated or calibrated using a two-step approach. The discount factor β and the coefficient of relative risk aversion σ are set using standard values: $\beta = 0.96$ and $\sigma \in [2, 6]$.³⁰ When there is only one financial asset, the rate of return on this asset, r , is typically set to 4%. Survival probabilities by age are exogenous and come from life tables. Tax rates are also set exogenously, around 40% for nonretirees and 0% for households that are retired. Some parameter values related to housing are also set outside the model—specifically, the transaction cost of selling a house ζ is usually set somewhere between 5% and 10%,³¹ and the down-payment constraint for home buyers ϕ is set to 5%, 10%, or 20%.³²

The depreciation rate on owned houses is set between 1% and 3%, on the basis of evidence from the Bureau of Economic Analysis. The remaining parameters—given σ , the parameter determining the elasticity of substitution between consumption and housing ρ , the expenditure share on consumption as compared with housing α , the home ownership premium to utility, the maximum sized rental house $\bar{h}$, the rental price per unit q , and the bequest function—are estimated or calibrated to match a set of facts about housing and capital in the aggregate or over the life cycle. Most authors add parameters or tweak their model specification until they decide the model fits the data along the margins they feel are most important.

The life-cycle profile and stochastic process for earnings are taken directly from the Panel Study of Income Dynamics (PSID) or from studies that use PSID data.³³ For example, earnings from the PSID in logarithmic form are regressed on age, age squared, marital status, household composition, and family fixed effects to obtain the deterministic life-cycle component. The residuals of this regression can be used to recover the stochastic component of labor income. Earning processes are estimated separately for different educational attainment groups, and earnings in retirement are the average of observed income in each education group.

The treatment of the process for house prices differs widely. Sometimes house prices are fixed; sometimes house prices follow some exogenous process set outside the model; and sometimes house prices are endogenously determined inside the model. When the path for house prices is set outside the model, many authors specify a process that

³⁰ Chambers et al. (2009a,c) and Fisher and Gervais (2011) depart from the utility function we describe and consider utility of the form $\left(\frac{\alpha}{1-\rho_1}\right)c^{1-\rho_1} + \left(\frac{1-\alpha}{1-\rho_2}\right)c^{1-\rho_2}$. Chambers et al. (2009a,c) set $\rho_1 = 3$ and $\rho_2 = 1$. They argue that this is necessary to match the income elasticity of owned housing in their data. Fisher and Gervais (2011) set $\rho_1 = 1$ and $\rho_2 = 2$ to ensure that expenditures on housing rise with the price of housing.

³¹ The widely cited article for this estimate is Gruber and Martin (2003). Some authors calibrate ζ to capture both monetary and nonpecuniary moving costs.

³² Occasionally this value is estimated on the basis of data on loan-to-value ratios of first-time home buyers.

³³ Storesletten et al. (2004), Cocco et al. (2005), and Heathcote et al. (2010) are standard references.

replicates the autocorrelation and variance of house price indices such as the Federal Housing Finance Authority or Case–Shiller–Weiss indices. Other authors try to replicate the patterns for house prices that can be derived from PSID data.³⁴ The benefit of the PSID approach is that it can deliver an estimate of the variance of prices of individual housing units. For example, [Flavin and Yamashita \(2002\)](#) construct real, annual housing returns using self-reported housing values from the PSID in adjacent years from 1968 to 1992. They estimate that the mean return to housing (including imputed rents net of maintenance) is 6.6%, with a standard deviation of 14.2%, and show the correlation of housing returns with stock, bond, and T-bill returns is almost zero.³⁵ It seems fair to say that the literature still lacks high-quality data to pin down the return correlation matrix between stocks, bonds, and individual houses.³⁶

The correlation of shocks to income and shocks to house prices plays a key role in portfolio decisions in many of these models. Many authors assume that there is zero correlation between individual house price returns and individual income realizations. On the basis of 1970–1992 PSID data, [Cocco \(2005\)](#) finds a 55% correlation between average house prices and the common part of household income. In his model, he assumes that house prices are perfectly positively correlated with the aggregate component of income and positively but imperfectly correlated with temporary labor income shocks. Using Swedish data for 2000–2007, [Vestman \(2012\)](#) estimates a correlation of 63% between the common component of income growth and national house price growth. The overall correlation with individual income growth is 14%. As a summary of commonly held views, (a) individual housing returns contain modest but important metropolitan-level and national components in addition to a large idiosyncratic component and (b) the national components of house prices and income are correlated with each other and with stock prices. We suspect that local labor income and local house prices are also positively correlated at a fine level of geography, such as the level of census tracts or ZIP codes, but the micro evidence is lacking.

³⁴ The literature has recognized that the PSID sample is small and that self-reported home values are noisy and possibly upward biased. For example, an increase in self-reported house prices across time may be due to quality improvements or pure appreciation. Nevertheless, some have argued that homeowners, on average, can guess the value of their house within a few percentage points. See [Goodman and Ittner \(1992\)](#), [Bucks and Pence \(2008\)](#), [Benítez-Silva et al. \(2010\)](#), and [Davis and Quintin \(2014\)](#).

³⁵ The imputed rent, or dividend, on housing is set equal to a proportion of the house value, where the proportion is equal to the real interest rate and the property tax rate multiplied by the marginal income tax rate. The short-term interest rate is fixed to 5%.

³⁶ Other studies measuring housing returns using microdata are those of [Piazzesi et al. \(2007\)](#), who improve on the method of [Flavin and Yamashita \(2002\)](#), and [Landvoigt et al. \(2013b\)](#), who estimate high-quality housing price appreciation rates for San Diego. [Favilukis et al. \(2011\)](#) discuss measurement issues and compare several approaches.

12.4.3 Major findings in the portfolio literature

The ultimate aim of the portfolio choice literature is to produce a model that can match life-cycle profiles for the incidence of home ownership, the incidence of stock market participation, total wealth, and the percentage of total wealth accounted for by housing, stock, and bond market wealth—that is, portfolio shares. The literature also increasingly seeks to match systematic differences in portfolio shares across groups of households, sorted by tenant status (owner vs. renter), age, income, and net worth. In the remainder of this section, we discuss the major findings in three strands of the literature: housing and consumption over the life cycle, the interplay of optimal housing holdings and optimal financial asset decisions, and mortgage choice. In this section, we study models that (a) assume house prices are constant, or are stochastic but determined outside the model—we call these models “partial equilibrium models”—and (b) do not explicitly study how government policy affects housing decisions. In [Section 12.6](#), we discuss similar models that endogenize house prices, while in [Section 12.7](#) we study the impact of tax and other policies on housing outcomes.

12.4.3.1 *Housing and consumption patterns over the life cycle*

12.4.3.1.1 Home ownership over the life cycle

An early body of literature studies how various segments of the housing market are connected in simple life-cycle models where housing comes in discrete and fixed sizes ([Stein, 1995](#); [Ortalo-Magné and Rady, 2006](#)). In such models, people move “up” the housing ladder—that is, they buy more expensive housing—as they age or become richer. These models do not try to match the life-cycle data to the same degree as the models we discuss in this section; rather, they are designed to study equilibrium differences in house prices across various segments of the housing market. We discuss these models in more detail in [Section 12.6.3](#).

The work of [Fernández-Villaverde and Krueger \(2011\)](#) is one of the earlier articles in the literature: although the article was published in 2011, it was last revised in 2001. Fernández-Villaverde and Krueger ask if a partial equilibrium life-cycle model with housing and idiosyncratic income and mortality risk can fit the life-cycle profiles of consumption spending (hump shaped), spending on durable goods (hump shaped), and housing and financial wealth. They treat durable goods and housing interchangeably. Durable goods provide service flows and act as the sole source of borrowing constraints. The main finding is that the interaction of borrowing constraints (our Equation [12.11](#)) and consumer durables produces young agents who accumulate durables early in life and increase nondurable spending and financial asset positions later in life.

[Yang \(2009\)](#) revisits the work of [Fernández-Villaverde and Krueger \(2011\)](#) and argues that consumption of housing first increases over the life cycle but fails to decline in old age, unlike nondurable spending (recall [Figure 12.8](#)). She adds housing transaction costs

to the framework of [Fernández-Villaverde and Krueger \(2011\)](#). Borrowing constraints are needed to explain the accumulation of housing early, and transaction costs are necessary to explain the slow downsizing of housing later in life.

In related work, [Nakajima and Telyukova \(2012\)](#) study the standard model's predictions for retired households. They show that retired homeowners spend down their wealth more slowly than renters. This arises in their model because retired homeowners cannot easily borrow against their house and because they want to stay in their current house as long as possible, rather than sell and move. Relative to the standard model, Nakajima and Telyukova model medical expenses explicitly in the budget constraint and allow for a consumption floor in the event of a large medical expense shock, as in [Hubbard et al. \(1995\)](#) and [Ameriks et al. \(2010\)](#).

[Halket and Vasudev \(2014\)](#) study the life-cycle patterns of home ownership and mobility across metropolitan statistical areas (MSAs). They show that the standard model, modified to allow for Lucas–Prescott-style job mobility ([Lucas and Prescott, 1974](#)), can account for the increase in home ownership, increase in wealth, and decline in mobility over the life cycle. The young move more frequently than the old because they rent. Uncertainty about family size can also affect the decision to move.

On the topic of home ownership over the life cycle, [Fisher and Gervais \(2011\)](#) ask why home ownership rates for young people declined substantially from 1980 through 2000. They show that the trend toward marrying later and the increase in income risk can account for almost the entirety of the decline. Unlike most of the articles in this literature, they allow for only three ages in their model (young, middle aged, old) and households transition stochastically between ages.

12.4.3.1.2 Housing collateral to smooth consumption

[Hurst and Stafford \(2004\)](#) and subsequent researchers study how households use home equity to smooth consumption. In their model, housing does not enter utility. They show that a combination of low liquid asset holdings and a bad income realization predicts borrowing against home equity, and the borrowing allows better consumption smoothing. They use PSID data to demonstrate that the mechanism they describe is present in the data.

In related work using Danish household panel data, [Ejarque and Leth-Petersen \(2008\)](#) demonstrate that new first-time home buyers, who have depleted their financial assets and have borrowed as much as possible to purchase their house, respond to income shocks by reducing consumption. They show that a relatively standard model, such as the one described earlier, can replicate this fact.

Providing further empirical evidence on the housing collateral channel, [Hryshko et al. \(2010\)](#) use data from the PSID linked to MSA geography to show that, after becoming disabled or losing a job, households tap into home equity to smooth consumption when home prices are rising. When house prices are falling, after disability or job loss

households reduce consumption. The study authors predict how the empirical results would change if down-payment constraints, adjustment costs, or the correlation of income and price growth were to change.

A related body of literature studies the response of consumption to house prices using microdata. [Campbell and Cocco \(2007\)](#) use UK Family Expenditure Survey data to estimate an elasticity of new consumption spending to exogenous changes in wealth as large as 1.7 for old households that own their home, but close to zero for young households that rent, after controlling for interest rates, household income, and demographics. Consumption responds to predictable changes in house prices, which is consistent with a housing collateral effect. Since it is predictable changes in aggregate and not regional house prices that seem to matter, the collateral effect operates at the aggregate level. This evidence is consistent with the results of [Lustig and Van Nieuwerburgh \(2010\)](#), discussed later, who find evidence for an aggregate housing collateral effect in US MSA-level consumption data.

[Li and Yao \(2007\)](#) also study the differential impact of exogenous changes to house prices on various groups of agents in the economy. An unanticipated positive increase in house prices benefits old homeowners because their remaining life span is short and they can afford more consumption. Despite earning a capital gain, young homeowners are worse off because they have a longer horizon and face higher borrowing costs for their housing. Renting households are strictly worse off after the shock. [Li and Yao \(2007\)](#) also show that the relationship between the uncertainty and volatility of house prices and the probability of home ownership is ambiguous and depends on the degree of household risk aversion.

[Bajari et al. \(2013\)](#) estimate the parameters of the standard life-cycle model presented earlier in this section using PSID data along with auxiliary data on mortgage interest rates and house prices. They simulate the model and compare simulated household responses to a negative house price shock and, separately, a negative income shock. They show that in response to a negative shock to house prices, households accumulate more housing later in life but do not change the basic shape of their life-cycle patterns of spending and saving. In contrast, in response to a negative and unanticipated income shock, households reduce both their housing demand and their consumption.

[Attanasio et al. \(2011\)](#) study the aggregate implications of the life-cycle model described earlier. They show that an unanticipated increase in the level of house prices leads to smaller housing units but not a decline in the home ownership rate, an increase in consumption of the old, and a decrease in consumption of the young. Relative to [Li and Yao \(2007\)](#), [Attanasio et al. \(2011\)](#) add the restriction that the home equity constraint only binds at the time of the purchase, such that households can have negative equity if house prices decline. The model is calibrated to English data.

Finally, [Kaplan and Violante \(2014\)](#) point out that the illiquidity of housing affects the propensity to consume out of fiscal stimulus payments. Many households have substantial illiquid housing wealth but limited liquid wealth. A fiscal transfer, which is an increase in liquid wealth, increases aggregate consumption by a much larger amount than would be

predicted from a model economy in which housing and liquid assets are aggregated together and are both considered to be liquid.

12.4.3.1.3 House price risk and demand for housing

[Han \(2008\)](#) studies housing demand when house prices are uncertain and volatile and housing incurs transaction costs. Han emphasizes that although housing is risky, driving down demand, current housing is a hedge against future housing demand shocks since price changes of housing units in the same market are correlated. A related argument was made by [Sinai and Souleles \(2005\)](#). Han uses a variant of the life-cycle framework to determine when, given (a) life-cycle profiles for income and tastes for housing, (b) the probability of moving to a new market with a different level for house prices, and (c) the inherent volatility of house prices in every market, the hedging demand for housing dominates its inherent risk. [Han \(2008\)](#) concludes that the impact of uncertainty of house prices depends on households' future plans. When households expect to increase their holdings of housing in the future, they buy a bigger home today in response to an increase in house price uncertainty. If, instead, households expect to downsize in the future, they reduce their holdings of housing today in response to an increase in house price uncertainty.

[Halket and Amior \(2013\)](#) study the relationship of housing risk and home ownership. They document that house price volatility is negatively correlated with home ownership rates and low loan-to-value ratios at the MSA level. They also show that house price volatility is high where house prices are high, because these areas are areas where land's share of home value is high. They find that a relatively standard model of housing predicts that home ownership rates are low in high house price areas. The model can also explain why loan-to-value ratios are low in areas where house prices are volatile.

12.4.3.2 *The role of housing in the financial portfolio*

The work of [Flavin and Yamashita \(2002\)](#) is among the first articles in the housing and portfolio-choice literature; it considers a simple mean-variance framework without labor income risk. In addition to their data-based contribution discussed earlier, [Flavin and Yamashita \(2002\)](#) highlight that young households who own housing have a highly levered position that should incentivize them to take less risk in the rest of their portfolio. This intuition qualitatively explains why young households hold fewer stocks. The article also proposes a resolution to an asset allocation puzzle. Standard portfolio theory prescribes that households combine the risk-free assets with the efficient portfolio of all risky assets and vary the holdings of each to suit their level of risk aversion. Financial planners, in contrast, advise clients to vary the proportion of risky stocks and bonds. In [Flavin and Yamashita \(2002\)](#), the latter is optimal since households are at a corner of zero in terms of their risk-free rate holdings.

Unlike [Flavin and Yamashita \(2002\)](#), who consider only the optimal portfolio of owner-occupiers and do not explicitly model life-cycle income and savings decisions of households, [Yao and Zhang \(2004\)](#) study how households optimally choose their

portfolio of financial assets using a life-cycle model like the one described earlier where households in the model choose whether to rent or own housing in each period. Agents in their model face the following trade-off: house prices are uncertain and volatile, leading households to want to have fewer stocks, but homeowners can use home equity as a buffer against income shocks, leading households to want to have more stocks. They show that renters and owners choose substantially different portfolios of financial assets, highlighting that conclusions drawn about optimal portfolio allocations over the life cycle from models that do not include a rental/own housing choice may be misleading. For example, when a household transitions from renting to owning in the model, the share of stocks in total wealth falls, but the share of stocks in liquid wealth increases. The reason is that the low correlation of stock and housing returns and the high equity risk premium make holding stocks relatively attractive.

The work of [Cocco \(2005\)](#) is similar to that of [Yao and Zhang \(2004\)](#), and some of the insights—for example, about what drives changes to equity participation over the life cycle—are similar. [Cocco \(2005\)](#) does not allow for a housing tenure decision (everyone owns housing), but includes a fixed cost of stock market participation. The model results in the finding that younger and lower-wealth home-owning households do not participate in the stock market because of the fixed cost of participation. Rather, their portfolios are heavily tilted toward real estate and are highly levered, similarly to the data. The large and risky housing investment makes participation in risky stock markets relatively unattractive in the model. [Cocco \(2005\)](#) shows that a small fixed participation cost in equity markets generates substantial rates of nonparticipation. As households age, leverage declines, stock market participation rates increase, and the share of liquid assets held as stocks rises, just like in the data. Thus, the article shows that stock market participation patterns in the data are less puzzling once the impact of housing is considered.

Like [Yao and Zhang \(2004\)](#), [Vestman \(2012\)](#) studies stock market participation rates for homeowners and renters, but unlike the existing literature, he considers Epstein–Zin preferences, such that the intertemporal elasticity of substitution is governed by a different parameter than the degree of risk aversion. [Vestman \(2012\)](#) also allows these preference parameters to vary across households. His model matches the hump-shaped life-cycle profile of home ownership in the data and generates a flatter and more accurate life-cycle profile of stock market participation than [Cocco \(2005\)](#). [Vestman \(2012\)](#) shows that households with low-risk aversion and high elasticity of intertemporal substitution save less, invest less of those savings in equity, are more likely to not participate in the stock market and are more likely to rent. Using panel data from Sweden, [Vestman \(2012\)](#) documents that stock market participation drops by one-fifth in the year of home purchase, and his model generates a similar sized decline.

12.4.3.2.1 Introducing geography

One important feature of housing most of the literature abstracts from is its spatial aspect. A household's two largest assets, human wealth and housing wealth, are intimately tied to

the fortunes of the location where that household lives and works. A large body of literature in urban economics, discussed elsewhere in this handbook, centers around issues of geography. For our purposes, we highlight a few articles that integrate spatial considerations into the financial portfolio choice literature. In [Ortalo-Magné and Prat \(2013\)](#) and [Hizmo \(2012\)](#), households choose where to live once at the beginning of life, and then in each subsequent period they decide on the portfolio composition.³⁷ For tractability, preferences exhibit constant absolute risk aversion, and consumption occurs at the end of life. In [Ortalo-Magné and Prat \(2013\)](#), households must pay rent where they work but they can invest in housing in every region, separating housing consumption from housing ownership. In [Hizmo \(2012\)](#), households own a home where they work. Markets are incomplete in that there are not enough risky assets to span all the shocks that hit labor income.

Focusing on from the results in [Hizmo \(2012\)](#), once the location has been determined, households optimally hold more of a stock whose returns hedge local income risk. Stocks whose returns covary strongly with house prices are poor hedges and require higher risk premiums. In addition to making risk sharing incomplete, the unspanned regional risk distorts the efficient spatial allocation of labor. Risk-averse households may end up in regions with lower house price volatility rather than where they will be most productive. The normative implication of this work is that households are willing to pay up handsomely for securities that reduce house price volatility.

12.4.3.3 Mortgage choice

[Campbell and Cocco \(2003\)](#) study the risk and return features of the typical mortgage contracts, FRMs or ARMs, when labor income, house prices, and real interest rates are uncertain and the size of the house is predetermined. The expectations hypothesis of the term structure is assumed to hold in their model: long-term bonds and FRMs contain no term risk premium. [Campbell and Cocco \(2003\)](#) note that ARMs are risky because payments might fluctuate more than income or might rise when incomes fall, forcing homeowners to reduce consumption. FRMs are expensive even when inflation is relatively stable because the cost of an FRM includes the value of a prepayment option—if interest rates fall, households have the option to refinance their FRM after paying a small monetary cost. FRMs are cheap only when inflation is high; the study authors refer to this as the wealth risk of FRMs. They document that households with large houses relative to income, with volatile labor income, and with high-risk aversion have a preference for an FRM. Households with a high probability of moving typically

³⁷ [Van Nieuwerburgh and Weill \(2010\)](#) and [Davis et al. \(2013\)](#) study spatial equilibrium models where agents optimally choose where to live and are mobile in each period. However, while both articles study the equilibrium interactions of housing and migration decisions, neither article studies a portfolio choice problem. [Han \(2013\)](#) also studies differences in expected returns to housing across metropolitan areas, but her study is largely empirical in focus.

prefer an ARM to take advantage of the lower rate. The study authors propose an inflation-indexed FRM, a contract that removes the income risk associated with ARMs but also the wealth risk associated with FRMs.

The share of households that finance a house with an FRM fluctuates significantly over time. To explain this phenomenon, [Kojien et al. \(2009\)](#) solve a two-period mortgage choice model where risk-averse households trade off the expected payments on an FRM and an ARM contract with the risk of these payments. The model generates an intuitive risk-return trade-off for mortgage choice: the ARM contract is more desirable the higher the nominal bond risk premium, the lower the variability of the real rate, and the higher the variability of expected inflation. The model predicts that time variation in the aggregate FRM share is caused by time variation in the bond risk premium, defined as the difference between the long-term bond yield (or FRM rate) and the expected average future short-term bond yield (ARM rate) over the life of the contract. [Kojien et al. \(2009\)](#) specify households form expectations over future short rates inside the model using vector autoregressions, blue chip forecaster data, or a backward-looking weighted average of past short rates. All three produce bond risk premiums whose fluctuations line up with the observed fluctuations in the ARM share in the United States, with the rule of thumb giving the strongest results. [Moench et al. \(2010\)](#) provide out-of-sample support for this theory by showing how the recent decline in rule-of-thumb bond risk premiums can help explain the unusually low ARM share between 2007 and 2010. [Badarinza et al. \(2013\)](#) extend the analysis of term-structure determinants of mortgage choice to multiple countries. [Campbell \(2013\)](#) studies mortgage market design around the world.

A natural question to ask is how the asset side (portfolio choice) interacts with the liability side (mortgage choice) of a household's balance sheet. [Van Hemert \(2010\)](#) extends the basic framework of [Cocco \(2005\)](#) and [Yao and Zhang \(2004\)](#) to include more interesting mortgage choices. He allows households to hold stocks, 1-, 3-, or 10-year long-term nominal bonds, and an ARM or an FRM. In his framework, ARMs are short positions in 1-year bonds, while FRMs are short positions in 10-year bonds. In the benchmark calibration, which features stochastic interest rates, households prefer to finance their house with an ARM to avoid paying the bond risk premium present in an FRM. Since young agents have most of their wealth in human capital, which is analogous to a bond, they invest their financial asset portfolio mostly in stocks. Middle-aged households hold some long-term bonds to hedge against real interest rate changes affecting their ARM. This long-term bond position increases as investors age and the value of their human capital declines. A risk-averse investor in retirement holds a negative position in 10-year bonds and a positive position in short-term bonds, similarly to a hybrid ARM. That position hedges real interest rate risk while avoiding exposure to inflation risk.

In sum, the portfolio choice and life-cycle literature have taken great steps toward understanding the economic forces motivating households' observed asset and liability

choices. Housing and mortgage choice are not only key components of households' asset and liability structure, but also play a crucial role in understanding household demand for financial assets and liabilities. In the next section, we investigate how households' demand for housing and financial assets affects equilibrium asset prices.

12.5. HOUSING AND ASSET PRICING

Given that housing importantly affects households' optimal portfolio choices, such as stock holdings, the presence of housing may affect how stocks and other assets are priced in equilibrium. In this section, we review the insights from the asset pricing literature with housing. Specifically, we study endowment economies and discuss models with production in the next section. It is well known that the canonical endowment economies of Lucas (1978) and Breeden (1979) with constant relative risk aversion preferences fail to match the asset pricing data. In particular, predicted risk premiums are too low and do not vary much over time and the risk-free rate is too high and too volatile (Hansen and Singleton, 1983; Mehra and Prescott, 1985; Weil, 1989). Adjusting these models to allow for production, as in Jermann (1998) and others, amplifies these problems.

The asset pricing literature has made great strides in the past two decades in determining the required properties of stochastic discount factors (SDFs) necessary to generate asset-pricing behavior that looks like the data. In summary, SDFs need to be persistent, countercyclical, and heteroskedastic—in particular, higher conditional variance is required in bad times. The external habit framework (Campbell and Cochrane, 1999), the long-run risk framework (Bansal and Yaron, 2004; Bansal et al., 2012), and the variable rare disaster framework (Gabaix, 2012) all deliver SDFs that have these properties, at least in endowment economies. Assumptions on preferences and technology in these frameworks are difficult to test directly. For this reason, economists studying housing have asked if SDFs can be generated that look like the data in a relatively standard model once observable housing-market constraints and frictions are considered.

12.5.1 Representative agent model

The baseline model extends the representative agent endowment economy of Lucas (1978) to allow for two distinct types of “trees” in the economy, one that yields nonhousing goods and services as “fruit” (dividends) and the other that yields housing services. The representative agent owns both types of trees and consumes all fruits. For simplicity, call the nonhousing goods and services nonhousing consumption, denoted c , and call housing services housing, denoted h . In any period t , households in this model choose consumption c_t , housing to purchase h_t at price per unit p_t , and the quantities of each of $i = 1, \dots, N$ assets A_{t+1}^i to maximize the expected net present value of utility

$$\sum_{t=0}^{\infty} \beta^t E[u(c_t, h_t)]$$

subject to the budget constraint in each period

$$c_t + p_t h_t + \sum_i^N A_{t+1}^i \leq \sum_i^N A_t^i \mathcal{R}_t^i + p_t h_{t-1} + w_t,$$

where w_t is income. h_{t-1} and $A_t^i \mathcal{R}_t^i$ for each i are predetermined as of the start of period t ; the future realizations of $\mathcal{R}_{t+1}^i$ may be random.

Denote the Lagrange multiplier on the budget constraint at period t as λ_t . The first-order conditions for consumption at t and for the optimal choice of period $t+1$ holdings of asset i are

$$\begin{aligned} c_t: \lambda_t &= \frac{\partial u}{\partial c_t}, \\ A_{t+1}^i: \lambda_t &= \beta E[\lambda_{t+1} \mathcal{R}_{t+1}^i]. \end{aligned}$$

Since households can freely purchase any asset, the second equation above must hold for all assets $i = 1, \dots, N$. When we combine the above first-order conditions, the return on any asset must satisfy the following equation:

$$1 = E_t[M_{t+1} \mathcal{R}_{t+1}^i]. \quad (12.12)$$

where M_{t+1} is β times the ratio of the marginal utility of consumption at $t+1$ to marginal utility of consumption at t . M_{t+1} is the SDF. One way to read Equation (12.12) is that it specifies that all assets must pay the same expected return after accounting for risk. The term that determines the required compensation for risk for each asset is the SDF; specifically, the required compensation for risk is largely determined by the covariance of M and $\mathcal{R}$.

Piazzesi et al. (2007) explore the asset-pricing implications of this model when households are assumed to have constant relative risk aversion preferences over a CES bundle of consumption and housing:

$$U(c_t, h_t) = \frac{\tilde{c}_t^{1-\sigma}}{1-\sigma}, \quad \tilde{c}_t = (\alpha c_t^\rho + (1-\alpha)h_t^\rho)^{\frac{1}{\rho}}, \quad (12.13)$$

where α is the weight on nonhousing consumption and $\varepsilon = \frac{1}{1-\rho}$ is the elasticity of substitution between c and h . The Cobb–Douglas case ($\varepsilon = 1$ or $\rho = 0$) is a special case and is discussed later.

Given preferences as specified in Equation (12.13), the logarithm of the SDF, call it m_{t+1} , can be written as the product of the standard single-good factor (nonhousing consumption growth) and a new factor that captures the effect of the composition of the bundle of consumption and housing:

$$m_{t+1} = \log \beta - \sigma \Delta \log c_{t+1} + \frac{1-\rho-\sigma}{\rho} \log \left(\frac{1+S_{t+1}}{1+S_t} \right) \quad \text{and} \quad S_t = \frac{c_t}{q_t h_t}. \quad (12.14)$$

q_t is defined as the rental price of one unit of housing, and thus S_t is the ratio of total nonhousing consumption to the rental value of housing.

In the special case of $\varepsilon = 1$ —that is, Cobb–Douglas preferences—optimal budget shares on consumption and housing are fixed and the ratio S_t is constant. In this case, the final term of the SDF in Equation (12.14) vanishes. The SDF reverts to that of the one-good endowment economy with its problematic asset-pricing predictions.

Piazzesi et al. (2007) consider values for ε strictly greater than, but close to 1—that is, ρ slightly positive. This choice makes the coefficient in front of the final term negative and large in absolute value. Given this parameterization, consider the covariance of the SDF and the return on an asset. An asset whose return is low when growth in the total expenditure to housing expenditure ratio $1 + S$ is low is risky—thus, it will require a high expected return. When ρ is slightly larger than zero, the two-factor model with housing has the potential to explain equity risk premiums. The model also implies that the ratio of nonhousing to housing consumption predicts future stock returns, a prediction Piazzesi et al. (2007) show is supported in the data.

Davis and Martin (2009) estimate the preference parameters generated by this model and argue that it fails to simultaneously price a portfolio of stocks and T-bills. Their generalized method of moments estimate of ρ is 0.2 (standard error 0.05). This value is too far from zero to deliver the required amplification to the standard Lucas–Breeden kernel in order to simultaneously price stocks and bonds. Davis and Martin (2009) expand their model to allow for leisure and home production rather than housing to enter utility. In each of these cases, they show that the model cannot price a portfolio of stocks and T-bills.

In related work, Yogo (2006) studies a model with Epstein–Zin preferences over a CES aggregate of nondurable and durable consumption. He shows that when the elasticity of substitution between nondurable and durable consumption is higher than the intertemporal elasticity of substitution, the marginal utility of consumption rises when durable consumption falls, which is in bad times. Using asset return data, he estimates a value of the intratemporal elasticity ε of 0.5–0.7 and a low value of 0.023 for the elasticity of substitution (alongside an unappealingly high value for the coefficient of relative risk aversion). If that 0.5–0.7 estimate for ε is valid for housing, too little action will be generated by the SDF to deliver plausible asset-pricing implications.

Like the previous articles, Flavin and Nakagawa (2008) consider a model with CES-aggregated preferences over nondurable consumption and durable housing. However, they study how the illiquidity of housing alters the SDF, building on the seminal article by Grossman and Laroque (1990). The presence of nonconvex adjustment costs makes the house a state variable, generating an SDF that displays considerable volatility and that depends on the history of wealth. Furthermore, adjustment costs allow a decoupling of

relative risk aversion from the intertemporal elasticity of substitution while maintaining standard preferences—that is, time-separable utility. Given the similar properties of the SDF, the housing model with adjustment costs produces a structural interpretation of an external habit. Furthermore, the housing adjustment model outperforms both the standard external habit model and the constant relative risk aversion model. The Euler equation for individual household's housing returns cannot be rejected, and the study authors estimate a plausible parameter value of $\sigma = 1.8$. They also estimate $\varepsilon = 0.13$ ($\rho = -6.7$), suggesting that housing expenditure shares rise by a lot in response to an increase in house prices, a result that is at odds with work by [Davis and Ortalo-Magné \(2011\)](#) and others that estimates that the expenditure share on housing is roughly fixed.

Finally, recent work by [Giglio et al. \(2014\)](#) uses data on housing, specifically data on long-term property leases, to reexamine the shape and structure of SDFs. [Giglio et al. \(2014\)](#) compare the prices of freeholds (indefinite ownership of property) with those of leaseholds (very long term but finite ownership) in England, Wales, and Singapore. They find that leaseholds trade at a substantial discount to freeholds, implying that housing services in the distant future are discounted at a surprisingly low rate. This result contradicts implications of the articles by [Campbell and Cochrane \(1999\)](#), [Bansal and Yaron \(2004\)](#), [Bansal et al. \(2012\)](#), and [Gabaix \(2012\)](#) discussed earlier. Combined with the high average returns on housing, inclusive of the entire stream of housing services, the work of [Giglio et al. \(2014\)](#) suggests that most of the reward for investing in housing reflects compensation for near-term risk rather than long-term risk, consistent with the findings of [van Binsbergen et al. \(2012\)](#) for the equity market. Their result suggests future researchers should search for preferences and constraints such that an asset-pricing model can generate a downward sloping term structure of housing and equity risk premiums.

12.5.2 Risk sharing with housing collateral

[Lustig and Van Nieuwerburgh \(2007\)](#) study the asset-pricing implications when housing is used as collateral. They model a heterogeneous agent economy with two “Lucas”-type trees, nonhousing goods and housing, as in the model of [Piazzesi et al. \(2007\)](#). The aggregate endowment of dividends of both trees is stochastic, and households differ in their realizations of the nonhousing endowment (labor income). Households have access to a full set of securities to share their income risk with each other. The key friction in the model is that households cannot commit to repaying their debt. This friction limits the degree of risk sharing that can be achieved as in the limited commitment model of [Alvarez and Jermann \(2000, 2001\)](#). Unlike other models in which households that default are denied access to financial markets, the punishment for default in [Lustig and Van Nieuwerburgh \(2007\)](#) is the loss of housing collateral. As a result, in equilibrium, households' borrowing is limited to the value of the collateral. Equivalently, the housing collateral constraint can be thought of as a solvency constraint which keeps households' net

worth strictly positive. Since future labor income cannot be pledged, housing is the only collateralizable asset in the model.

A key state variable of the model is the ratio of housing wealth to total wealth, the “housing collateral” ratio. When the housing collateral ratio is high, risk sharing is nearly complete and the economy’s allocations and prices are close to those of the representative-agent Lucas economy. When housing collateral is scarce, risk sharing is incomplete. In this environment, agents who have received persistent positive income shocks require an increase in their share of nonhousing and housing consumption to encourage them to continue to participate in risk-sharing arrangements. Unconstrained agents experience a reduction in their share of aggregate consumption. The net effect is that cross-sectional distribution of consumption growth widens and risk sharing deteriorates.

Lustig and Van Nieuwerburgh (2007) show that the logarithm of the SDF of this economy contains a new term, $\sigma \Delta \log \xi_{t+1}^a$, which measures the extent to which housing collateral constraints bind in the economy³⁸:

$$m_{t+1} = \log \beta - \sigma \Delta \log C_{t+1} + \frac{1 - \rho - \sigma}{\rho} \log \left(\frac{1 + S_{t+1}}{1 + S_t} \right) + \sigma \Delta \log \xi_{t+1}^a. \quad (12.15)$$

Note that the housing collateral effect operates even when preferences are separable between housing and nonhousing consumption, or when the aggregator of these two goods in utility is of the Cobb–Douglas type. A key implication of this framework is that the degree of risk sharing and equilibrium asset prices vary with the housing collateral ratio. The persistent and countercyclical movements in the housing collateral ratio lead to persistent countercyclical movements in the SDF. Times of scarce housing collateral lead to high market prices of risk and high conditional volatility (heteroskedasticity) in the SDF. In addition, the model delivers a downward sloping term structure of equity and housing risk premiums, consistent with the findings of van Binsbergen et al. (2012) and Giglio et al. (2014) cited above.

Lustig and Van Nieuwerburgh (2005) test three asset-pricing predictions of the model of Lustig and Van Nieuwerburgh (2007). First, a low housing collateral ratio should predict periods of high market prices of risk and therefore high future excess returns on stocks. Regressions on the predictability of US stock returns confirm this. Second, in the cross section, risky assets have returns that covary strongly with aggregate consumption growth when housing collateral is scarce and opportunities for risk sharing are limited, as predicted by the model. Third, as shown in Equation (12.15), the model predicts that an augmented “conditional” consumption capital asset pricing

³⁸ Specifically, ξ_{t+1}^a is a cross-sectional moment of individual ξ_{t+1}^i . The latter are cumulative Lagrange multipliers on the housing collateral constraint. These multipliers increase over time whenever an agent’s constraint binds, but otherwise stay constant. When no agent’s constraint binds, $\Delta \log \xi_{t+1}^a = 0$.

model, that is, a consumption capital asset pricing model augmented with a housing collateral term, should be able to fit data on returns. Using the housing-collateral consumption capital asset pricing model augmented implied by Equation (12.14), the model is able to account for more than 80% of the cross-sectional variation in size, book-to-market portfolios, a long-term bond portfolio, and the overall stock market portfolio. It is also able to reconcile the difference in expected returns between value and growth stocks.³⁹

An advantage of the limited commitment framework is that it gives rise to a unique SDF despite the presence of endogenously incomplete markets. Because unconstrained agents price the assets at each date and state of the world, the SDF is volatile, which is key for asset-pricing predictions. A disadvantage of the model is that default only occurs for “strategic” reasons. In the data, many households default for liquidity reasons—that is, low income realizations. The next section discusses models of housing default in some detail.

12.6. THE HOUSING BOOM AND BUST AND THE GREAT RECESSION

A recent body of literature explores models that can simultaneously generate plausible business cycle moments, as discussed in Section 12.3, realistic life-cycle consumption-savings and portfolio profiles, as discussed in Section 12.4, and sufficiently volatile house and asset prices, as discussed in Section 12.5. Much of the work in this area focuses on the epic housing boom and bust of the first decade of the twenty-first century and studies how changes in the lending environment affected home ownership rates, house prices, and other macroeconomic aggregates. These articles usually depart from the representative agent framework: agents differ by age, income, and wealth; some agents rent and some own housing; and some agents borrow and others lend money. We start by discussing several articles that assume that house prices are fixed—either they are set outside the model or they are trivially pinned down by a simple production function for housing. These articles differ from the portfolio literature discussed earlier in that they study events of the past 15 years and focus on heterogeneity across agents. We then review studies where house prices are determined endogenously in the model economy. In both cases, we distinguish between models that allow for default and those that do not. The models with default naturally focus on the housing bust. While these models are simpler to solve, the literature with exogenous house prices faces the criticism that it ignores the fact that changes in the model environment associated with the housing boom or bust may well affect house prices.

³⁹ In addition, Lustig and Van Nieuwerburgh (2010) document using quantity data that the degree of risk sharing between US metropolitan areas decreases when housing collateral is scarce.

12.6.1 Exogenous house prices, no default

A first branch of the literature studies the consequences of innovation in housing finance, through either lower down-payment constraints or the availability of new mortgage contracts. [Silos \(2007a\)](#) studies an equilibrium model with owner-occupied housing where agents differ in age, income, and wealth and the price of housing is fixed. He shows that a relaxation of down-payment constraints does not change the business-cycle properties of the model, but affects residential and nonresidential investment decisions of the young and poor. [Silos \(2007b\)](#) shows that adding rental housing as a choice greatly helps models such as those of [Fernández-Villaverde and Krueger \(2011\)](#) and [Diaz and Luengo-Prado \(2010\)](#) explain the distribution of wealth by age, since the choices of renters, who are younger and poorer on average, are not well captured by models without renting.

[Chambers et al. \(2009a\)](#) study a model with different mortgage choices to determine the extent to which reductions in down-payment constraints and availability of second mortgages increased the home ownership rate in the United States between 1994 and 2005. Agents in their model face idiosyncratic income and mortality risk, and idiosyncratic capital gains upon house sales but no aggregate risk. The model also includes a market for rental services. The study authors show that most of the increase in home ownership rates over this period is attributable to the availability of second mortgages. In closely related work, [Chambers et al. \(2009c\)](#) study how the multiple mortgage contracts that coexist in equilibrium affect home ownership rates, the size of owned housing, and risk sharing. They first assume the economy has one type of mortgage, the FRM, and then ask what happens if a second type of mortgage is added to the economy. Different types of mortgages are considered in the experiment. The addition of mortgage contracts that allow for lower down payments or increasing payment schedules over time increases the home ownership rate.

[Iacoviello and Pavan \(2013\)](#) study the business cycle properties of a life-cycle model with housing where agents differ in their discount factor and preference for renting. House prices are fixed and mortgage default is not allowed. They show that the model can replicate the basic life-cycle facts discussed earlier, but can also account for the procyclicality of household debt. The heterogeneity in the model is sufficiently rich to match observed inequality in wealth. The model attributes the Great Moderation (a period of reduced volatility of major macroeconomic aggregates that occurred from the early 1980s until the early years of the twenty-first century) to lower down-payment constraints and increased individual earnings volatility. On the one hand, the reduction in down payments leads to an increase in home ownership and a reduction in the volatility of housing investment because homeowners face adjustment costs in changing investment, whereas renters do not. On the other hand, the increase in earnings volatility makes people less likely to be homeowners and less willing to buy an asset subject to transactions costs. Thus, the impact on housing volatility is ambiguous, since renters adjust more but

homeowners adjust by much less, again due to transactions costs. The study authors argue that the reduction of housing investment during the Great Recession was the consequence of tightening financial conditions coupled with a bad aggregate productivity shock.

12.6.2 Exogenous house prices with default

A second branch of the literature aims to understand observed mortgage defaults in the housing bust by studying optimal mortgage default models that take house prices as given. House price shocks, unemployment shocks, home equity extraction, and new mortgage contracts are four (interacting) channels of interest in this literature.

Campbell and Cocco (2012) study how default rates vary by mortgage type (FRM, ARM, and interest-only mortgage). They solve the problem of a household that must decide how much to consume and whether to default on the mortgage. The household faces idiosyncratic income risk, as well as house price and interest rate risk. When home equity becomes sufficiently negative, households default. The negative-equity threshold for default depends on the degree to which households are borrowing constrained, which itself depends on income shocks, interest rates, and the terms of the mortgage contract (ARM vs. FRM). Higher loan-to-value ratios affect default probabilities by increasing the likelihood of negative equity and reducing incentives for repayment. A higher debt-to-income ratio also increases default by reducing the affordability of mortgages, making borrowing constraints more likely to bind and reducing the threshold level of negative equity that triggers default.

Taking house prices, interest rates, and aggregate income as given, Chen et al. (2013) investigate why households increased mortgage debt during the recent housing boom. In their model, agents can either rent or own housing, and if they own housing, they can finance it with a mortgage. In every period, agents can invest liquid assets at the risk-free rate, refinance a mortgage, take out a home equity loan (home equity line of credit), sell the home, or default. Both mortgages and home equity lines of credits are subject to loan-to-value and loan-to-income limits; mortgages are long-term contracts that can be refinanced at a cost. The article avoids the simplifying assumption made in most of the portfolio literature that mortgages are negative bond positions. Under the set of housing-finance frictions considered, the study authors demonstrate that the model can replicate the observed quantity of mortgage debt accumulation (Figure 12.7) given observed dynamics for income and house prices.

Like Chen et al. (2013), Laufer (2013) studies the importance of home equity extraction in accounting for the observed surge in mortgage defaults during the housing bust. In his model, households can refinance their mortgage at a cost and are subject to collateral constraints. Mortgages are nonrecourse in the model, but on default households are forced to rent. He estimates the parameters of his model using income data from the

PSID, asset data from the SCF, and data on the experiences of a large panel of Los Angeles homeowners. Given realized house-price dynamics, the model can replicate the observed time series of home equity extraction, home sales, and mortgage defaults at different loan-to-value ratios. In his model, homeowners extract equity when house prices increase and when liquid assets are close to zero. He shows that when homeowners have strongly positive house price expectations, they use home equity extraction to finance additional consumption. The model attributes the increase in mortgage defaults to negative house price shocks wiping out home equity rather than to negative income/unemployment spells. [Laufer \(2013\)](#) runs counterfactual experiments with his model to study the implications of two policies that might reduce default: tighter borrowing constraints and stronger recourse. Under both policies, defaults fall sharply and less home equity is extracted.

[Hatchondo et al. \(2013\)](#) add mortgage default to the standard life-cycle model described in the previous section. Similarly to what is done by [Campbell and Cocco \(2003\)](#), the size of housing is fixed. The study authors allow households in each period to choose their down payment (which then implies a particular fixed mortgage payment) and they then solve for the zero-profit mortgage rate associated with each mortgage.⁴⁰ Similar to [Li and Yao \(2007\)](#), Hatchondo et al. demonstrate that house price shocks are not an important source of consumption inequality. The focus of the article is on two policies that can reduce default: tighter borrowing constraints and stronger recourse, as in [Laufer \(2013\)](#). First, requiring a 15% down payment relative to an economy without a down-payment constraint minimally impacts the home ownership rate (it falls from 63.1% to 62.9%), but lowers defaults on mortgages by 30% (from 0.6% to 0.4%). Young potential home buyers are worse off because of the increase in down-payment constraints, but existing homeowners benefit since they can refinance at the lower interest rate that arises as a result of lower default. Second, a policy of temporary income garnishment as a punishment for default holds fixed the variance of consumption but otherwise reduces default rates, boosts home ownership rates, and reduces down-payment percentages.

[Corbae and Quintin \(2015\)](#) study the rise in foreclosures in the housing bust and ask how much can be explained by the large number of high-leverage mortgage contracts. They model households who choose between different mortgage contracts and face exogenous income and house price shocks. They show that the relaxation of credit constraints during the housing boom can account for more than 60% of the increase in foreclosures during the housing bust. This result holds despite the fact that each mortgage contract is priced such that mortgage originators earn zero expected profits, so the surge in foreclosures does not represent *ex ante* mispricing of loans. The result arises for two reasons: The higher loan-to-value ratios allowed after relaxation of constraints imply that

⁴⁰ Since households with positive equity can refinance, they can essentially choose their sequence of mortgage payments.

households are more likely to have negative equity in the event of a bad realization of house prices; and the relaxation of constraints allowed borrowers more prone to default (i.e., they had lower income and less assets) to purchase housing.

In related work, [Garriga and Schlagenhauf \(2009\)](#) argue that an essential feature to understand the spike in the foreclosure rate is leverage. An increase in leverage exposes homeowners to additional risk in the event of declines in house prices. To test the quantitative importance of the leverage channel, the study authors develop an equilibrium model of long-term mortgage choice and default. The model captures the pattern of foreclosure rates across loan products observed in the subprime crisis. The decline in house prices can account for most of the observed increase in the foreclosure rate and decline in home ownership in the United States.

In sum, the combination of home equity extraction during the boom—facilitated by new mortgage contracts—and sustained by rising house prices, combined with a negative house price shock during the bust, goes a long way toward accounting for the observed run-up of mortgage debt and the subsequent foreclosure rates. The natural next step in this literature is to endogenize house prices in the boom and bust.

12.6.3 Endogenous house prices, no default

Predating the housing boom and bust, [Stein \(1995\)](#) and [Ortalo-Magné and Rady \(2006\)](#) studied the effect of down-payment constraints on equilibrium house prices and housing transactions. They explain how changes to constraints or income for people purchasing at one end of the housing ladder alter the entire distribution of house prices. The last few years has seen a burst of research activity trying to account for the massive rise in housing prices during the boom in models with down-payment constraints.

Recently, [Kiyotaki et al. \(2011\)](#) use a general equilibrium life-cycle model to study the implications of an unexpected increase in land's share of housing in an environment where interest rates are set outside the model. They show that when land's share of the value of housing rises, land and house prices become more sensitive to productivity shocks and to shocks to world interest rates, causing a large redistribution of wealth between net buyers and sellers of housing in response to these shocks. They show that after an increase in the share of housing attributable to land, a tightening of financial constraints does not impact house prices.

In contrast with the previous article, [Chu \(2014\)](#) shows that a relaxation of credit constraints can cause a big change in the purchase price of owned houses relative to the rental price of rental properties if owner-occupied and rental properties are inelastically supplied and if conversion from rental property to owner-occupied property and vice versa is costly. [Chu \(2014\)](#) also finds that changes in supply of housing and changes in the process for income (an increase in the level and the volatility) importantly affected house prices from 1995 to 2005. He shows that the transition path from an initial steady state to a new

steady state at different levels of income volatility and credit constraints produces an “overshooting” of house prices—house prices initially rise above their new steady-state value and then slowly decline, as predicted by [Ortalo-Magné and Rady \(2006\)](#).

In earlier work, [Favilukis et al. \(2011\)](#) aim to explain why house prices are more volatile than rents. They also seek to explain the slow but large increase in the price–rent ratio during the housing boom. They study a two-sector general equilibrium model of housing and nonhousing production, using a production framework similar to that of [Davis and Heathcote \(2005\)](#), but where heterogeneous households face limited risk-sharing opportunities as a result of incomplete financial markets. A house in the model is a residential durable asset that provides utility to the household, is illiquid (expensive to trade), and can be used as collateral in debt obligations. The model economy is populated by a large number of overlapping generations of households who receive utility from both housing and nonhousing consumption and who face a stochastic life-cycle earnings profile. Market incompleteness arises because heterogeneous agents face idiosyncratic and aggregate risks against which they cannot perfectly insure themselves, and because of collateralized borrowing constraints on households, as discussed in [Section 12.4](#).

[Favilukis et al. \(2011\)](#) study the macroeconomic consequences of three systemic changes in housing finance, with an emphasis on how these factors affect risk premiums in housing markets, and how risk premiums in turn affect home prices: the impact of changes in housing collateral requirements, the change in borrowing costs (the spread of mortgage rates over risk-free debt), and the impact of an influx of foreign capital into the domestic bond market.⁴¹ These are meant to capture important changes to the US economy after the year 2000.⁴² To model capital inflows, the third structural change in the model, [Favilukis et al.](#) introduce foreign demand for a domestic risk-free bond into the market clearing condition for that asset. Foreign capital purchases of the risk-free US bond are determined outside the model.⁴³

According to the model, price–rent ratios increased during the housing boom owing to the simultaneous occurrence of positive aggregate productivity shocks and a relaxation of credit standards. Both of these events generated an endogenous decline in risk premiums on housing and equity assets, and the decline in housing risk premiums generated the increase in house prices relative to (imputed) rents. Risk premiums in the model fell for two reasons. First, lower collateral requirements directly increased access to credit,

⁴¹ [Garriga et al. \(2012\)](#) also study the impacts of these changes on land and house prices, but they assume a representative agent and do not allow for aggregate shocks to productivity.

⁴² [Van Nieuwerburgh \(2012\)](#) and [Favilukis et al. \(2013\)](#) provide detailed evidence on all three changes and supporting references.

⁴³ [Krishnamurthy and Vissing-Jorgensen \(2012\)](#) estimate that such foreign governmental holders, such as central banks, have a zero price elasticity for US Treasuries, because they are motivated by reserve currency or regulatory motives ([Kohn, 2002](#)).

which acted as a buffer against unexpected income declines. Second, lower costs of borrowing reduced the expense of obtaining the collateral required to increase borrowing capacity and provide insurance. The model attributes the housing bust to a sudden tightening of credit constraints and a set of negative economic shocks.

Favilukis et al. (2011) attribute changes in price–rent ratios during the housing boom and bust largely to changes in housing risk premiums and not to changes in risk-free interest rates. In the model, if credit standards become laxer, the need for precautionary savings falls, which, by itself, generates an increase in risk-free interest rates. If price–rent ratios increase following a relaxation of credit standards, absent other changes, the model suggests that housing risk premiums must have declined by more than the increase in risk-free rates.

Of course, during the housing boom, risk-free interest rates did not increase, but rather declined. Favilukis et al. attribute the decline in risk-free rates to an increase in foreign purchases of domestic bonds. They suggest that the decline in interest rates did not, by itself, cause house prices to boom relative to rents. The reason is that foreign purchases of domestic bonds forced US savers to purchase more equity and housing than they would have otherwise desired, thus increasing risk premiums on housing and lowering house prices and the price of other risky assets.⁴⁴ In addition, Favilukis et al. suggest that foreign capital inflows stimulated residential investment, raising the expected stock of future housing and pushing down house prices. They thus suggest the net effect of a large capital inflow from abroad into safe securities depressed real risk-free interest rates but had only a small effect on house prices.

In summary, Favilukis et al. argue two opposing forces affected the price of housing risk during the housing boom and bust. During the boom, credit standards were relaxed, lowering risk premiums, and foreigners bought more US bonds, raising risk premiums. With the model's calibrated parameters, the decline in risk premiums from the effects of the former exceeded the rise in risk premiums from the latter. During the bust, in which capital inflows held constant but credit standards were tightened to their preboom levels, risk-free interest rates remained low but risk premiums rose, lowering house prices relative to rents.⁴⁵

Boldrin et al. (2013) explore the role of the construction sector in explaining changes to aggregate employment and output during the housing boom. They emphasize that the construction sector has important interlinkages with other sectors of the economy, and that variation in the demand for residential investment propagates to aggregate output through these linkages. After calibrating these linkages using input–output data, the study

⁴⁴ Campbell et al. (2009) show that prior to the housing boom and bust, the historical covariance of risk-free interest rates and the risk premium to housing was negative.

⁴⁵ In related work, Favilukis et al. (2012) study the welfare implications of a reversal of (stochastic) foreign purchases of safe US debt.

authors estimate that the construction sector accounted for 29% of the growth in employment and 8% of the change in GDP during 2002–2007. They also estimate the contribution of construction to the decline in employment was 28% and its contribution to the decline of GDP was 43% during the Great Recession.

In a detailed study of the housing market of San Diego, [Landvoigt et al. \(2013b\)](#) emphasize heterogeneity in the quality of housing. They show that the areas of San Diego that experienced the largest housing boom also experienced the greatest bust. They propose a model in which households are assigned to houses of various quality, and study how changes in income and asset distributions changed assignments between 2000 and 2005. [Landvoigt et al. \(2013b\)](#) show that a relaxation of credit constraints played an important role in determining the cross-sectional patterns of capital gain to house prices. The study complements the previous literature in that it infers the importance of the relaxation of credit constraints on house prices using data from within one metropolitan area. It also extends the housing ladder model of [Ortalo-Magné and Rady \(2006\)](#) by considering more quality types of housing.

12.6.4 Endogenous house prices with default

A final strand of the general equilibrium heterogeneous agent literature studies mortgage default. These models add endogenous house price determination to the partial equilibrium framework of the default models discussed in [Section 12.6.2](#). The feedback between foreclosures and house prices is important in accounting for the substantial house price decline in the bust. Particularly, [Chatterjee and Eyigungor \(2009, 2011, 2012\)](#) and [Hedlund \(2014\)](#) build general equilibrium models of housing in order to evaluate the effects of the drop in house prices and a change in housing supply on equilibrium foreclosure rates. Like the default models with exogenous house prices, their focus is to understand the underlying reasons for, and consequences of, the recent foreclosure crisis.

[Chatterjee and Eyigungor \(2009\)](#) attribute the foreclosure crisis to overbuilding: an increase in housing supply not matched by an increase in demand. House prices must fall to absorb the excess supply. Because of leverage, the initial fall in house prices pushes some households into foreclosure and forces them to demand a smaller house in the rental market. For housing markets to clear, house prices must adjust downward even further.

[Jeske et al. \(2013\)](#) assume mortgage contracts last one period and study the effects of a specific government housing market policy, as discussed in [Section 12.7](#). Combining the assumption that contracts last one period with perfect competition delivers a sharp characterization of equilibrium mortgage interest rates and default policies.⁴⁶ Minimum down-payment requirements arise endogenously in the model of [Jeske et al. \(2013\)](#). The same result occurs in [Arslan et al. \(2013\)](#), who study a model with default and

⁴⁶ In related work, [Mitman \(2012\)](#) considers the interaction of recourse and bankruptcy on the decision to default in an environment with one-period mortgages and costless refinancing.

endogenous house prices and evaluate how the model's predictions change when down-payment constraints, interest rates, or unemployment rates change.

Hedlund (2014) models a search friction and emphasizes that housing illiquidity, measured by the probability of selling a home (or the time on the market), increases the probability of default for a financially distressed homeowner. Mortgage banks charge a higher default risk premium on new mortgages in times of housing illiquidity, which tightens borrowing constraints and causes mortgage illiquidity. Higher mortgage rates force some households to sell rather than to refinance their mortgage. These tend to be households with high loan-to-value loans. With high asking prices, their homes are unlikely to sell, worsening housing liquidity. More homeowners also go into foreclosure, and the selling delays associated with real-estate-owned properties further increase housing illiquidity. The interaction between housing and mortgage illiquidity increases equilibrium house price volatility. It generates house prices, mortgage debt, time on the market, and foreclosure dynamics in line with the data. A policy that improves lenders' recourse reduces house price and residential investment volatility, increases existing sales volatility, and all but eliminates foreclosures.

In sum, the literature has made great strides in accounting for the boom and bust in house prices in models with heterogeneous borrowers and lenders, and borrowing constraints that depend on endogenously determined house prices. In response to large and unforeseen changes in credit constraints, these models can generate booms and busts in house prices of nearly the same amplitude as observed in the 2000–2010 period. While this result is useful and informative, it leaves unanswered a set of primary questions that in our view should be the focus of future research: Why did credit constraints change, why were the changes unforeseen, and are large and unforeseen changes in credit constraints necessary to generate large house price fluctuations or can other changes or mechanisms generate the same-sized booms and busts? The models in this section also do not allow any direct feedback from changes in house prices and defaults to measured aggregate productivity and economic activity more broadly. Exploring this link seems quite important as casual empiricism suggests the Great Recession was caused by a financial crisis resulting from a relatively small number of mortgage defaults.

12.7. HOUSING POLICY

A host of government policies subsidize mortgage debt and provide financial incentives for home ownership. Economists and policy makers are concerned with the welfare implications of these policies and their impact on home ownership, house prices, mortgage debt, and financial stability. In this section, we briefly discuss research that evaluates the effect of housing policy on outcomes, such as housing demand and house prices, and welfare using the tools of modern quantitative macroeconomics. In our view, this is an important topic where research is needed. Events of the past 15 years suggest that the

existing housing finance architecture has proven deficient in providing a stable environment in which the young or poor would be comfortable with the risks associated with home ownership. Studying the effects of a major overhaul of that system requires a general equilibrium analysis, using the tools described in this chapter. We close by discussing briefly literature studying the interaction of monetary policy and housing/mortgage markets, and recent literature evaluating mortgage modification programs.

12.7.1 Mortgage interest rate deductibility

The early literature focuses on the effects of the mortgage interest rate deductibility. One of the earliest articles in this literature is by [Gervais \(2002\)](#), who studies the impact of the tax code on the accumulation of housing and other wealth in a general equilibrium model. He finds that the failure to tax imputed rents from owner-occupied housing and the tax deductibility of mortgage interest distort the rate of return on housing capital as compared with business capital. If the government were to tax imputed rents, holding total revenue collected unchanged by lowering income taxes, the stock of business capital would increase by 6%, the stock of housing capital would decrease by 8%, and one-quarter of households that own housing would switch to renting. The elimination of mortgage interest deductibility leaves the total stock of housing capital unchanged, but increases the rental stock relative to the owner-occupied stock. Both policies would be uniformly welfare improving. Note that [Gervais \(2002\)](#) holds house prices fixed—output can be costlessly transformed at a one-to-one rate into consumption, business investment, or residential investment.

[Chambers et al. \(2009b\)](#) document that most rental properties in the United States are owned by households. They model the decision to invest in rental housing and analyze the connection between the asymmetric tax treatment of homeowners and landlords and the progressivity of income taxation. They find that eliminating the mortgage interest deduction—assuming budget neutrality—has a positive effect on home ownership because it lowers the average tax rate in the economy. This leads to an increase in average household income and wealth. Under the assumption that house prices are fixed, the increase in wealth prompts a switch from renting to owning.

In contrast to the previous articles, [Sommer and Sullivan \(2013\)](#) endogenize not only rental but also ownership prices. They include a very detailed treatment of the tax code in their model. They show that repealing tax deductions causes house prices to decline but does not affect rents. The decline in house prices and price-rent ratios boosts home ownership as it encourages more young people to save for a down payment on a house. In some simulations of alternative government tax policies, the home ownership rate rises by almost eight percentage points. Eliminating mortgage interest rate or property tax deductibility also shifts housing consumption from the rich to the poor and is associated with an overall welfare gain.

Floetotto et al. (2012) investigate the same policy changes but focus on the welfare effects in the transition between steady states. Upon removal of the tax on rental income and the interest rate deductibility of mortgage interest, a quarter of households suffer initial welfare losses, driven by house prices that fall by 4% on impact before recovering to a 1.6% decline relative to the initial steady state. The initial welfare losses are largely borne by middle-income earners in their model. The alternative policy of eliminating the asymmetric tax treatment of owner-occupied and rental housing by taxing imputed rents leads to a larger fraction of initial losers (a third) and affects wealthy households the most. This article underscores the importance of endogenizing house prices when evaluating how tax policy affects housing markets. In sum, abolishing current tax expenditures that benefit homeowners would substantially increase welfare, increase home ownership, and reduce inequality, but may lead to temporary capital losses for existing homeowners.

12.7.2 Housing finance intermediaries

A conceptually similar question asks how the government-sponsored enterprises (GSEs) Fannie Mae and Freddie Mac distort US housing markets.⁴⁷ Jeske et al. (2013) study the effect of bailout guarantees to the GSEs, modeled as a 0.3 percentage point subsidy to the mortgage interest rate and financed by taxes on income. This interest rate subsidy is regressive, hurting low-income, low-asset renters and homeowners with a small mortgage, while benefiting the wealthy. The subsidy affects the amount of leverage in the housing system and its distribution, but has little effect on the home ownership rate. Eliminating the subsidy increases aggregate welfare and is especially beneficial to low-income, low-asset households.

Studying the GSEs is part of an important area of future research that will take seriously the role that intermediaries play in the housing finance system. The financial regulatory reforms enacted in the Great Recession prompted economists to understand better how regulation affects the behavior of these financial intermediaries and the housing outcomes they influence. Quantitative research on this important topic has just begun. Landvoigt (2012) models a banking sector that intermediates mortgage credit between borrowers and savers, and studies the effects of bank capital regulation and the cost of raising equity on house prices, risk sharing, and welfare. He also studies the effect of lower cost mortgage securitization as an alternative to mortgage banking. He concludes that securitization accounted for approximately 30% of the increase in mortgage debt during the period in which collateral constraints on housing were relaxed. In the model, securitization lowers borrowing costs, leading to more valuable collateral; and the lower intermediation costs cause a faster transition to a high-debt regime.

⁴⁷ See Acharya et al. (2011) for a discussion of the historical evolution of the GSEs and a plan to reform them.

12.7.3 Housing and monetary policy

A chapter on housing and the macroeconomy would be incomplete without mentioning literature that explores the role that housing plays in transmitting monetary policy. The two best-known articles in this area are by [Iacoviello \(2005\)](#) and [Iacoviello and Neri \(2010\)](#). The production side of [Iacoviello and Neri \(2010\)](#) is similar to that of [Davis and Heathcote \(2005\)](#), but other key features of the model are different. Rather than a representative household, [Iacoviello and Neri \(2010\)](#) have two types of households, patient and impatient, with impatient households always borrowing constrained. An increase in house prices allows impatient households to borrow and consume more. The model also has features common to the new-Keynesian framework: monopolistically competitive firms, sticky prices, and a monetary authority setting interest rates according to a Taylor rule. The benefit of this additional richness is that the model can match many business cycle moments. [Iacoviello and Neri \(2010\)](#) demonstrate that monetary policy shocks account for about 15–20% of the cyclical variation in house prices and residential investment. They also show that technology shocks and monetary policy shocks cannot fully account for the increase in house prices during the housing boom. Rather, their model attributes two-thirds of the increase in house prices during the boom to shocks to preferences for housing.⁴⁸

Recent work by [Garriga et al. \(2013\)](#), building on the work of [Kydland et al. \(2012\)](#), explores how monetary policy can affect the dynamics of residential investment. Unlike the articles by [Iacoviello \(2005\)](#) and [Iacoviello and Neri \(2010\)](#), they specify mortgage payments are in nominal terms and do not model any other nominal rigidity. Changes in monetary policy affect the distribution of real payments of mortgages over the life of the loan (the “price” effect) but also affect the amount of interest paid (the “wealth” effect). The study authors show that monetary policy has a great influence on housing in economies with ARMs rather than FRMs: price and wealth effects tend to reinforce each other with ARMs and offset each other with FRMs.

[Midrigan and Philippon \(2011\)](#) examine the cross-sectional relationship between household leverage, house prices, and unemployment, motivated by the empirical evidence in [Mian and Sufi \(2009\)](#). They observe that the regions with the largest changes to household debt also experienced the largest declines in employment and output during the financial crisis. To match this fact, they study a model with a cash-in-advance constraint, but where households can borrow against their home equity using a home equity line of credit. They show that the presence of cash-in-advance constraints generates a decline in employment after borrowing constraints are tightened: the tightening of liquidity constraints reduces the velocity of money and triggers a recession. To account

⁴⁸ [Aruoba et al. \(2014\)](#) also study the impact of monetary policy on housing in a model where housing is a capital input to home production. Their focus is on the relationship between steady-state inflation rates and the level of house prices and aggregate welfare.

for the cross-sectional pattern of leverage and employment in the data, the model also features wage rigidities and labor reallocation frictions.

Feroli et al. (2012) argue that a physical overhang of existing homes, depressed house prices due to foreclosures, and tight credit conditions all impaired the normal monetary transmission mechanism. They argue that policies that subsidize the refinancing of underwater mortgages and speed the transition from foreclosure to real estate owned by banks will improve the effectiveness of monetary policy.

12.7.4 Mortgage modification programs

Related literature analyzes the 2009 Housing Affordable Modification Program (HAMP) and asks how many foreclosures HAMP prevented and at what cost (Agarwal et al., 2012; Scharlemanny and Shorez, 2013; Hembre, 2014). We think a potentially interesting line of research embeds HAMP and other foreclosure-relief policies into a general equilibrium model and studies the extent to which these policies altered the level of house prices and overall welfare.

12.8. CONCLUSION

The recent housing boom and bust and financial crisis have renewed the curiosity of economists of all stripes about the interplay of housing, finance, and macroeconomics. In this chapter, we have surveyed the state of research in this field with an emphasis on its development over the past decade. While much progress has been made, many challenges and interesting problems remain.

One unsettled debate concerns the origins of the housing boom and bust and, related, why a relatively modest number of mortgage defaults precipitated a financial crisis and a severe recession. A common explanation is that housing boomed owing to the availability of exotic mortgage products and an expansion of mortgage credit. This pushes the question back one level. Specifically, why did these mortgage products become available or more widely adopted, and did the availability of these products cause house prices to rise, or did the products emerge because market participants expected continued appreciation of housing? Similarly, if house price expectations are to blame, what triggered the change in expectations?⁴⁹

In our view, researchers should be focusing on the role played by housing and real estate in causing financial crises, if any, and on the welfare consequences of government policies designed to encourage home ownership and reduce foreclosures. In addition, more work also remains to be done in integrating the local aspects of housing and labor markets into standard portfolio choice, asset pricing, and macroeconomic models with housing.

⁴⁹ For a discussion of these issues, see Gerardi et al. (2008), Glaeser et al. (2013), Van Nieuwerburgh (2012), and Foote et al. (2012).

ACKNOWLEDGMENTS

We thank Gilles Duranton, Jack Favilukis, Carlos Garriga, David Kohn, Ralph Koijen, Sydney Ludvigson, Erwan Quintin, and Will Strange for comments and suggestions, and Vadim Elenev for excellent research assistance.

REFERENCES

- Acharya, V.V., Richardson, M., Van Nieuwerburgh, S., White, L.J., 2011. *Guaranteed to Fail: Freddie, Fannie, and the Debacle of Mortgage Finance*. Princeton University Press, Princeton, NJ.
- Agarwal, S., Amromin, G., Ben-David, I., Chomsisengphet, S., Piskorski, T., Seru, A., 2012. Policy intervention in debt renegotiation: evidence from the home affordable modification program. Fisher College of Business, Working paper No. 2012-03-020.
- Albrecht, J., Anderson, A., Smith, E., Vroman, S., 2007. Opportunistic matching in the housing market. *Int. Econ. Rev.* 48 (2), 641–664.
- Albrecht, J., Gautier, P.A., Vroman, S., 2010. Directed search in the housing market. Georgetown University, Working paper.
- Alvarez, F., Jermann, U.J., 2000. Efficiency, equilibrium, and asset pricing with risk of default. *Econometrica* 68 (4), 775–798.
- Alvarez, F., Jermann, U.J., 2001. Quantitative asset pricing implications of endogenous solvency constraints. *Rev. Financ. Stud.* 14, 1117–1152.
- Ameriks, J., Caplin, A., Laufer, S., Van Nieuwerburgh, S., 2010. The joy of giving or assisted lining? Using strategic surveys to separate bequest and precautionary motives. *J. Financ.* 66 (2), 519–561.
- Arslan, Y., Guler, B., Taskin, T., 2013. Joint dynamics of house prices and foreclosures. Indiana University, Working paper.
- Aruoba, S.B., Davis, M.A., Wright, R., 2014. Homework in monetary economics: inflation, home production, and the production of homes. University of Wisconsin-Madison, Working paper.
- Attanasio, O.P., Bottazzi, R., Low, H.W., Nesheim, L., Wakefield, M., 2011. Inflation, nominal debt, housing and welfare. Institute for Fiscal Studies, Working paper.
- Badarinar, C., Campbell, J.Y., Ramadoraim, T., 2013. What calls to ARMs? International evidence on interest rates and the choice of adjustable-rate mortgages. Harvard University, Working paper.
- Bajari, P., Chan, P., Krueger, D., Miller, D., 2013. A dynamic model of housing demand: estimation and policy implications. *Int. Econ. Rev.* 54 (2), 409–442.
- Bansal, R., Yaron, A., 2004. Risks for the long run: a potential resolution of asset pricing puzzles. *J. Financ.* 59 (4), 1481–1509.
- Bansal, R., Kiku, D., Yaron, A., 2012. An empirical evaluation of the long-run risks model for asset prices. *Crit. Financ. Rev.* 1 (1), 183–221.
- Benhabib, J., Rogerson, R., Wright, R., 1991. Homework in macroeconomics: household production and aggregate fluctuations. *J. Polit. Econ.* 99 (6), 1166–1187.
- Benitez-Silva, H., Heiland, F., Jimenez-Martin, S., 2010. How well do individuals predict the selling prices of their home? SUNY-Stony Brook, Working paper.
- Boldrin, M., Garriga, C., Peralta-Alva, A., Sanchez, J.M., 2013. Reconstructing the great recession. Federal Reserve Bank of St. Louis, Working paper 2013-006B.
- Breedon, D.T., 1979. An intertemporal asset pricing model with stochastic consumption and investment opportunities. *J. Financ. Econ.* 7 (3), 265–296.
- Bucks, B., Pence, K., 2008. Do homeowners know their house values and mortgage terms? *J. Urban Econ.* 64 (2), 218–233.
- Burns, A., Mitchell, W., 1946. *Measuring Business Cycles*. NBER, New York.
- Burnside, C., Eichenbaum, M., Rebelo, S., 2011. Understanding booms and busts in housing markets. NBER Working paper 16734.
- Campbell, J.Y., 2013. Mortgage market design. *Rev. Financ.* 17 (1), 1–33.

- Campbell, J.Y., Cocco, J.F., 2003. Household risk management and optimal mortgage choice. *Q. J. Econ.* 118 (4), 1449–1494.
- Campbell, J.Y., Cocco, J.F., 2007. How do house prices affect consumption? Evidence from micro data. *J. Monet. Econ.* 54 (3), 591–621.
- Campbell, J.Y., Cocco, J.F., 2012. A model of mortgage default. Harvard University, Working paper.
- Campbell, J.Y., Cochrane, J.H., 1999. By force of habit: a consumption-based explanation of aggregate stock market behavior. *J. Polit. Econ.* 107 (2), 205–251.
- Campbell, S.D., Davis, M.A., Gallin, J., Martin, R.F., 2009. What moves housing markets? A variance decomposition of the rent-price ratio. *J. Urban Econ.* 66 (2), 90–102.
- Case, K.E., Quigley, J.M., Shiller, R.J., 2005. Comparing wealth effects: the stock market versus the housing market. *Adv. Macroecon.* 5 (1), 1–32.
- Catte, P., Girouard, N., Price, R.W., Andre, C., 2004. Housing markets, wealth and the business cycle. OECD Economics Department, Working papers No. 394.
- Chambers, M., Garriga, C., Schlagenhauf, D.E., 2009a. Accounting for changes in the homeownership rate. *Int. Econ. Rev.* 50 (3), 677–726.
- Chambers, M., Garriga, C., Schlagenhauf, D.E., 2009b. Housing policy and the progressivity of income taxation. *J. Monet. Econ.* 56 (8), 1116–1134.
- Chambers, M., Garriga, C., Schlagenhauf, D.E., 2009c. The loan structure and housing tenure decisions in an equilibrium model of mortgage choice. *Rev. Econ. Dyn.* 12 (3), 444–468.
- Chatterjee, S., Eyigungor, B., 2009. Foreclosures and house price dynamics: a quantitative analysis of the mortgage crisis and the foreclosure prevention policy. Federal Reserve Bank of Philadelphia, Working paper 09–22.
- Chatterjee, S., Eyigungor, B., 2011. A quantitative analysis of the US housing and mortgage markets and the foreclosure crisis. Federal Reserve Bank of Philadelphia, Working paper 11–26.
- Chatterjee, S., Eyigungor, B., 2012. Maturity, indebtedness, and default risk. *Am. Econ. Rev.* 102 (6), 2674–2699.
- Chen, H., Michaux, M., Roussanov, N., 2013. Houses as ATMs? Mortgage refinancing and macroeconomic uncertainty. University of Southern California, Working paper.
- Chu, Y., 2014. Credit constraints, inelastic supply, and the housing boom. *Rev. Econ. Dyn.* 17 (1), 52–69.
- Cocco, J.F., 2005. Portfolio choice in the presence of housing. *Rev. Financ. Stud.* 18 (2), 535–567.
- Cocco, J.F., Gomes, F.J., Maenhout, P.J., 2005. Consumption and portfolio choice over the life cycle. *Rev. Financ. Stud.* 18 (2), 491–533.
- Cooper, D., Dynan, K., 2013. Wealth shocks and macroeconomic dynamics. Federal Reserve Bank of Boston, Working paper.
- Corbae, D., Quintin, E., 2015. Leverage and the foreclosure crisis. *J. Polit. Econ.*, forthcoming.
- Corradin, S., Fillat, J.L., Vergara-Alert, C., 2014. Optimal portfolio choice with predictability in house prices and transaction costs. *Rev. Financ. Stud.* 27 (3), 823–880.
- Davis, M.A., Heathcote, J., 2005. Housing and the business cycle. *Int. Econ. Rev.* 46 (3), 751–784.
- Davis, M.A., Heathcote, J., 2007. The price and quantity of residential land in the United States. *J. Monet. Econ.* 54 (8), 2595–2620.
- Davis, M.A., Martin, R.F., 2009. Housing, home production, and the equity and value premium puzzles. *J. Hous. Econ.* 18 (2), 81–91.
- Davis, M.A., Ortalo-Magné, F., 2011. Household expenditures, wages, rents. *Rev. Econ. Dyn.* 14 (2), 248–261.
- Davis, M.A., Palumbo, M., 2001. A primer on the economics and time series econometrics of wealth effects. Federal Reserve Board of Governors, Working paper FEDS 2001–09.
- Davis, M.A., Quintin, E., 2014. Default when house prices are uncertain. University of Wisconsin–Madison, Working paper.
- Davis, M.A., Lehnert, A., Martin, R.F., 2008. The rent-price ratio for the aggregate stock of owner-occupied housing. *Rev. Income Wealth* 54 (2), 279–284.
- Davis, M.A., Fisher, J.D.M., Veracierto, M., 2013. Gross migration, housing and urban population dynamics. Federal Reserve Bank of Chicago, Working paper 2013–19.

- Davis, M.A., Fisher, J.D.M., Whited, T., 2014. Macroeconomic implications of agglomeration. *Econometrica* 82 (2), 731–764.
- Diaz, A., Luengo-Prado, M.J., 2010. The wealth distribution with durable goods. *Int. Econ. Rev.* 51 (1), 143–170.
- Dorofeenko, V., Lee, G.S., Salyer, K.D., 2014. Risk shocks and housing supply: a quantitative analysis. *J. Econ. Dyn. Control* 45, 194–219.
- Ejarque, J., Leth-Petersen, S., 2008. Consumption and savings of first time owners: how do they deal with adverse income shocks. University of Copenhagen, Working paper.
- Favilukis, J., Ludvigson, S.C., Van Nieuwerburgh, S., 2011. The macroeconomic effects of housing wealth, housing finance and limited risk sharing in general equilibrium. Unpublished Paper, New York University.
- Favilukis, J., Ludvigson, S.C., Van Nieuwerburgh, S., 2012. Foreign ownership of U.S. safe assets: good or bad? Unpublished Paper, New York University.
- Favilukis, J., Kohn, D., Ludvigson, S.C., Van Nieuwerburgh, S., 2013. International capital flows and house prices: theory and evidence. In: Glaeser, E.L., Sinai, T. (Eds.), *Housing and the Financial Crisis*. National Bureau of Economic Research, pp. 235–299.
- Fernández-Villaverde, J., Krueger, D., 2011. Consumption and saving over the life cycle: how important are consumer durables. *Macroecon. Dyn.* 15 (5), 725–770.
- Feroli, M., Harris, E., Sufi, A., West, K., 2012. Housing, monetary policy, and the recovery. In: *Proceedings of the U.S. Monetary Policy Forum 2012*.
- Fisher, J.D.M., 1997. Relative prices, complementarities and comovement among components of aggregate expenditures. *J. Monet. Econ.* 39 (3), 449–474.
- Fisher, J.D.M., 2007. Why does household investment lead business investment over the business cycle. *J. Polit. Econ.* 115 (1), 141–168.
- Fisher, J.D.M., Gervais, M., 2011. Why has home ownership fallen among the young. *Int. Econ. Rev.* 52 (3), 883–912.
- Flavin, M., Nakagawa, S., 2008. A model of housing in the presence of adjustment costs: a structural interpretation of habit persistence. *Am. Econ. Rev.* 98 (1), 474–495.
- Flavin, M., Yamashita, T., 2002. Owner-occupied housing and the composition of the household portfolio. *Am. Econ. Rev.* 92 (1), 345–362.
- Floetotto, M., Kirker, M., Stroebel, J., 2012. Government intervention in the housing market: who wins, who loses? Stanford University, Working paper.
- Foote, C.L., Gerardi, K.S., Willen, P.S., 2012. Why did so many people make so many ex post bad decisions? The causes of the foreclosure crisis. FRB Boston Public Policy Discussion Paper Series, Paper No. 12–2.
- Gabaix, X., 2012. Variable rare disasters: an exactly solved framework for ten puzzles in macro finance. *Q. J. Econ* 127 (2), 645–700, Working paper NYU Stern.
- Garriga, C., Schlagenhauf, D.E., 2009. Home equity, foreclosures, and bailouts. Federal Reserve Bank of St. Louis, Working paper.
- Garriga, C., Manuelli, R., Peralta-Alva, A., 2012. A model of price swings in the housing market. Federal Reserve Bank of St. Louis, Working paper 2012-022A.
- Garriga, C., Kydland, F.E., Sustek, R., 2013. Mortgages and monetary policy. NBER Working paper 19744.
- Garriga, C., Chambers, M., Schlagenhauf, D.E., 2014. Did housing policies cause the post-war housing boom? A general equilibrium analysis. In: Fishback, P., Snowden, K., White, E. (Eds.), *Housing and Mortgage Markets in Historical Perspective*. University of Chicago Press.
- Gerardi, K., Lehnert, A., Sherlund, S.M., Willen, P., 2008. Making sense of the subprime crisis. *Brookings Papers on Economic Activity*, pp. 69–145.
- Gervais, M., 2002. Housing taxation and capital accumulation. *J. Monet. Econ.* 49 (7), 1461–1489.
- Ghent, A.C., Owyang, M.T., 2010. Is housing the business cycle? Evidence from U.S. cities. *J. Urban. Econ.* 67 (3), 336–351.
- Ghysels, E., Plazzi, A., Torous, W., Valkanov, R., 2013. Forecasting real estate prices. In: Elliott, G., Timmermann, A. (Eds.), *Handbook of Economic Forecasting*, vol. II. Elsevier, pp. 509–580.

- Giglio, S., Maggiori, M., Stroebe, J., 2014. Very long-run discount rates. New York University, Working paper.
- Glaeser, E.L., Gottlieb, J.D., Gyourko, J., 2013. Can cheap credit explain the housing boom? In: Glaeser, E.L., Sinai, T. (Eds.), *Housing and the Financial Crisis*. University of Chicago Press, pp. 301–359.
- Gomme, P., Kydland, F.E., Rupert, P., 2001. Home production meets time to build. *J. Polit. Econ.* 109 (5), 1115–1131.
- Goodman Jr., J.L., Ittner, J.B., 1992. The accuracy of home owners' estimates of house value. *J. Hous. Econ.* 2 (4), 339–357.
- Green, R.K., 1997. Follow the leader: how changes in residential and non-residential investment predict changes in GDP. *Real Estate Econ.* 25 (2), 253–270.
- Greenwood, J., Hercowitz, Z., 1991. The allocation of capital and time over the business cycle. *J. Polit. Econ.* 99 (6), 1188–1214.
- Greenwood, J., Rogerson, R., Wright, R., 1995. Household production in real business cycle theory. In: Cooley, T.F. (Ed.), *Frontiers of Business Cycle Research*. Princeton University Press, pp. 157–174.
- Grenadier, S.R., 1995. The persistence of real estate cycles. *J. Real Estate Financ. Econ.* 10 (2), 95–119.
- Grossman, G., Laroque, G., 1990. Asset pricing and optimal portfolio choice in the presence of illiquid durable consumption goods. *Econometrica* 58 (1), 25–51.
- Gruber, J., Martin, R.F., 2003. Precautionary savings and the wealth distribution with illiquid durables. Federal Reserve Board of Governors, Working paper.
- Halket, J., Amior, M., 2013. Do households use homeownership to insure themselves? Evidence across U.S. cities. University of Essex, Working paper.
- Halket, J., Vasudev, S., 2014. Saving up or settling down: home ownership over the life cycle. *Rev. Econ. Dyn.* 17 (2), 345–366.
- Han, L., 2008. Hedging house price risk in the presence of lumpy transaction costs. *J. Urban Econ.* 64 (2), 270–287.
- Han, L., 2013. Understanding the puzzling risk-return relationship to housing. *Rev. Financ. Stud.* 26 (4), 877–928.
- Hansen, L.P., Singleton, K., 1983. Stochastic consumption, risk aversion, and the temporal behavior of asset returns. *J. Polit. Econ.* 91 (2), 249–265.
- Hatchondo, J.C., Martinez, L., Sanchez, J.M., 2013. Mortgage defaults. Federal Reserve Bank of St. Louis, Working paper.
- He, C., Wright, R., Zhu, Y., 2013. Housing and liquidity. University of Wisconsin-Madison, Working paper.
- Head, A., Lloyd-Ellis, H., 2012. Housing liquidity, mobility, and the labour market. *Rev. Econ. Stud.* 79 (4), 1559–1589.
- Head, A., Lloyd-Ellis, H., Sun, H., 2011. Search and the dynamics of house prices and construction. Queens University, Working paper.
- Heathcote, J., Storesletten, K., Violante, G.L., 2010. The macroeconomic implications of rising wage inequality in the United States. *J. Polit. Econ.* 118 (4), 681–722.
- Hedlund, A., 2014. The cyclical dynamics of illiquid housing, debt, and foreclosures. University of Missouri, Working paper.
- Hembre, E., 2014. HAMP, home attachment, and mortgage default. University of Wisconsin-Madison, Working paper.
- Hirata, H., Kose, M.A., Otrok, C., Terrones, M.E., 2013. Global house price fluctuations: synchronization and determinants. In: Giavazzi, F., West, K.D. (Eds.), *NBER International Seminar on Macroeconomics 2012*. National Bureau of Economic Research, pp. 119–166.
- Hizmo, A., 2012. Risk in housing markets: an equilibrium approach. New York University, Working paper.
- Hodrick, R.J., Prescott, E.C., 1997. Postwar U.S. business cycles: an empirical investigation. *J. Money Credit Bank* 29 (1), 1–16.
- Hornstein, A., Praschnik, J., 1997. Intermediate inputs and sectoral comovement in the business cycle. *J. Monet. Econ.* 40 (3), 573–595.
- Hryshko, D., Luengo-Prado, M.J., Sorenson, B.E., 2010. House prices and risk sharing. *J. Monet. Econ.* 57 (8), 975–987.

- Hubbard, G.R., Skinner, J., Zeldes, S.P., 1995. Precautionary saving and social insurance. *J. Polit. Econ.* 103 (2), 360–399.
- Hurst, E., Stafford, F., 2004. Home is where the equity is: mortgage refinancing and household consumption. *J. Money Credit Bank.* 36 (6), 985–1014.
- Iacoviello, M., 2005. House prices, borrowing constraints, and monetary policy in the business cycle. *Am. Econ. Rev.* 95 (3), 739–764.
- Iacoviello, M., Neri, S., 2010. Housing market spillovers: evidence from an estimated DSGE model. *Am. Econ. J. Macroecon.* 2 (2), 125–164.
- Iacoviello, M., Pavan, M., 2013. Housing and debt over the life cycle and over the business cycle. *J. Monet. Econ.* 60 (2), 221–238.
- Jermann, U.J., 1998. Asset pricing in production economies. *J. Monet. Econ.* 41 (2), 257–275.
- Jeske, K., Krueger, D., Mitman, K., 2013. Housing, mortgage bailout guarantees and the macro economy. *J. Monet. Econ.* 60 (8), 917–935.
- Kahn, J.A., 2008. What drives housing prices. Federal Reserve Bank of New York, Staff Report No. 345.
- Kaldor, N., 1957. A model of economic growth. *Econ. J.* 67 (268), 591–624.
- Kaplan, G., Violante, G.L., 2014. A model of the consumption response to fiscal stimulus payments. *Econometrica* 82 (4), 1199–1239.
- Keys, B.J., Mukherjee, T., Seru, A., Vig, V., 2009. Financial regulation and securitization: evidence from subprime loans. *J. Monet. Econ.* 56 (5), 700–720.
- Keys, B.J., Mukherjee, T., Seru, A., Vig, V., 2010. Did securitization lead to lax screening? Evidence from subprime loans. *Q. J. Econ.* 125 (1), 307–362.
- Keys, B., Piskorski, T., Seru, A., Vig, V., 2013. Mortgage financing in the housing boom and bust. In: Glaeser, E., Sinai, T. (Eds.), *NBER Book Housing and the Financial Crisis*. pp. 143–204.
- Keys, B.J., Seru, A., Vig, V., 2012. Lender screening and the role of securitization: evidence from prime and subprime mortgage markets. *Rev. Financ. Stud.* 25 (7), 2071–2108.
- Kiyotaki, N., Michaelides, A., Nikolov, K., 2011. Winners and losers in housing markets. *J. Money Credit Bank.* 43 (2–3), 255–296.
- Kohn, D., 2002. Panel: implications of declining treasury debt. What should the federal reserve do as treasury debt is repaid? *J. Money Credit Bank.* 34 (3), 941–945.
- Koijen, R.S.J., Hemert, O.V., Van Nieuwerburgh, S., 2009. Mortgage timing. *J. Financ. Econ.* 93 (2), 292–324.
- Krishnamurthy, A., Vissing-Jorgensen, A., 2012. The aggregate demand for treasury debt. *J. Polit. Econ.* 120 (2), 233–267.
- Kydland, F.E., Prescott, E.C., 1982. Time to build and aggregate fluctuations. *Econometrica* 50 (6), 1345–1370.
- Kydland, F.E., Rupert, P., Sustek, R., 2012. Housing dynamics over the business cycle. NBER, Working paper 18432.
- Landvoigt, T., 2012. Aggregate implications of the increase in securitized mortgage debt. Stanford University, Working paper.
- Landvoigt, T., Piazzesi, M., Schneider, M., 2013a. Housing assignment with restrictions: theory and evidence from the Stanford campus. Stanford University, Working paper.
- Landvoigt, T., Piazzesi, M., Schneider, M., 2013b. The housing market(s) of San Diego. Stanford University, Working paper.
- Laufer, S., 2013. Equity extraction and mortgage default. Federal Reserve Board, Working paper.
- Leamer, E.E., 2007. Housing is the business cycle. NBER Working paper No. 13428.
- Li, W., Yao, R., 2007. The life-cycle effects of house price changes. *J. Money Credit Bank.* 39 (6), 1375–1409.
- Lucas, R.E.J., 1978. Asset prices in an exchange economy. *Econometrica* 46 (6), 1429–1454.
- Lucas, R.E.J., Prescott, E.C., 1974. Equilibrium search and unemployment. *J. Econ. Theory* 7 (2), 188–209.
- Lustig, H., Van Nieuwerburgh, S., 2005. Housing collateral, consumption insurance and risk premia: an empirical perspective. *J. Financ.* 60 (3), 1167–1219.
- Lustig, H., Van Nieuwerburgh, S., 2007. Can housing collateral explain long-run swings in asset returns? New York University, Working paper.

- Lustig, H., Van Nieuwerburgh, S., 2010. How much does housing collateral constrain regional risk sharing? *Rev. Econ. Dyn.* 13 (2), 265–294.
- Mankiw, N.G., Weil, D.N., 1989. The baby boom, the baby bust, and the housing market. *Reg. Sci. Urban Econ.* 19 (2), 235–258.
- McGrattan, E.R., Rogerson, R., Wright, R., 1997. An equilibrium model of the business cycle with household production and fiscal policy. *Int. Econ. Rev.* 38 (2), 267–290.
- Mehra, R., Prescott, E.C., 1985. The equity premium: a puzzle. *J. Monet. Econ.* 15 (2), 145–161.
- Mian, A., Sufi, A., 2009. The consequences of mortgage expansion: evidence from the U.S. mortgage default crisis. *Q. J. Econ.* 124 (4), 1449–1496.
- Mian, A., Sufi, A., 2011. House prices, home equity-based borrowing, and the U.S. household leverage crisis. *Am. Econ. Rev.* 101 (5), 2132–2156.
- Mian, A., Sufi, A., 2012. The effects of fiscal stimulus: evidence from the 2009 cash for clunkers program. *Q. J. Econ.* 127 (3), 1107–1142.
- Mian, A., Sufi, A., 2014. What explains the 2007–2009 drop in employment? University of Chicago, **Working paper**.
- Mian, A., Sufi, A., Trebbi, F., 2010. The political economy of the US mortgage default crisis. *Am. Econ. Rev.* 100 (5), 1967–1998.
- Mian, A., Rao, K., Sufi, A., 2013. Household balance sheets, consumption, and the economic slump. *Q. J. Econ.* 128 (4), 1687–1726.
- Mian, A., Sufi, A., Trebbi, F., 2014. Foreclosures, house prices, and the real economy. University of Chicago, **Working paper**.
- Midrigan, V., Philippon, T., 2011. Household leverage and the recession. New York University, **Working paper**.
- Mitman, K., 2012. Macroeconomic effects of bankruptcy and foreclosure policies. University of Pennsylvania, **Working paper**.
- Moench, E., Vickery, J.I., Aragon, D., 2010. Why is the market share of adjustable-rate mortgages so low? *Curr. Issues Econ. Financ.* 16 (8), 1–11.
- Muellbauer, J.N., Murphy, A., 1997. Booms and busts in the UK housing market. *Econ. J.* 107 (445), 1701–1727.
- Nakajima, M., Telyukova, I.A., 2012. Home equity in retirement. Federal Reserve Bank of Philadelphia, **Working paper**.
- Ngai, L.R., Tenreyro, S., 2014. Hot and cold seasons in the housing market. *Am. Econ. Rev.* 104 (12), 3991–4026.
- Ortalo-Magné, F., Prat, A., 2013. Spatial asset pricing: a first step. Columbia University, **Working paper**.
- Ortalo-Magné, F., Rady, S., 2006. Housing market dynamics: on the contribution of income shocks and credit constraints. *Rev. Econ. Stud.* 73 (2), 459–485.
- Piazzesi, M., Schneider, M., 2009. Momentum traders in the housing market: survey evidence and a search model. *Am. Econ. Rev.* 99 (2), 406–411.
- Piazzesi, M., Schneider, M., Tuzel, S., 2007. Housing, consumption and asset pricing. *J. Financ. Econ.* 83 (3), 531–569.
- Piazzesi, M., Schneider, M., Stroebl, J., 2013. Segmented housing search. Stanford University, **Working paper**.
- Piskorski, T., Seru, A., Vig, V., 2010. Securitization and distressed loan renegotiation: evidence from the subprime mortgage crisis. *J. Financ. Econ.* 97 (3), 369–397.
- Scharlemanny, T.C., Shorez, S.H., 2013. Does reducing ‘underwater’ prevent mortgage default? Evidence from HAMP PRA. Georgia State University, **Working paper**.
- Silos, P., 2007a. Housing, portfolio choice and the macroeconomy. *J. Econ. Dyn. Control.* 31 (8), 2774–2801.
- Silos, P., 2007b. Housing tenure and wealth distribution in life cycle economies. *B.E. J. Macroecon* 7 (1), 1–24.
- Sinai, T., Souleles, N., 2005. Owner-occupied housing as a hedge against rent risk. *Q. J. Econ.* 120 (2), 763–789.

- Sommer, K., Sullivan, P., 2013. Implications of U.S. tax policy for house prices, rents and homeownership. Federal Reserve Board of Governors, Working paper.
- Stein, J., 1995. Prices and trading volume in the housing market: a model with down-payment effects. *Q. J. Econ.* 110 (2), 379–406.
- Storesletten, K., Telmer, C.I., Yaron, A., 2004. Consumption and risk sharing over the life cycle. *J. Monet. Econ.* 51 (3), 609–633.
- Topel, R., Rosen, S., 1988. Housing investment in the United States. *J. Polit. Econ.* 96 (4), 718–740.
- van Binsbergen, J., Brandt, M., Koijen, R., 2012. On the timing and pricing of dividends. *Am. Econ. Rev.* 102 (4), 1596–1618.
- Van Hemert, O., 2010. Household interest rate risk management. *Real Estate Econ.* 38 (3), 467–505.
- Van Nieuwerburgh, S., 2012. The research agenda: Stijn Van Nieuwerburgh on housing and the macroeconomy. *Econ. Dyn. Newslett.* 13(2). <http://www.EconomicDynamics.org/News261.htm#agenda>.
- Van Nieuwerburgh, S., Weill, P.-O., 2010. Why has house price dispersion gone up? *Rev. Econ. Stud.* 77 (4), 1567–1606.
- Vestman, R., 2012. Limited stock market participation among renters and home owners. Stockholm University, Working paper.
- Vissing-Jorgensen, A., 2002. Limited asset market participation and the elasticity of intertemporal substitution. *J. Polit. Econ.* 110 (4), 825–853.
- Weil, P., 1989. The equity premium puzzle and the risk-free rate puzzle. *J. Monet. Econ.* 24 (3), 401–421.
- Wong, Y.-Y., Wright, R., 2011. Buyers, sellers and middlemen: variations on search-theoretic themes. National Bureau of Economic Research, Working paper No. 17511.
- Yang, F., 2009. Consumption over the life cycle: how different is housing. *Rev. Econ. Dyn.* 12 (3), 423–443.
- Yao, R., Zhang, H.H., 2004. Optimal consumption and portfolio choices with risky housing and borrowing constraint. *Rev. Financ. Stud.* 18 (1), 197–239.
- Yogo, M., 2006. A consumption-based explanation of expected stock returns. *J. Financ.* 61 (2), 539–580.

This page intentionally left blank

CHAPTER 13

The Microstructure of Housing Markets: Search, Bargaining, and Brokerage

Lu Han, William C. Strange

Rotman School of Management, University of Toronto, Toronto, ON, Canada

Contents

13.1. Introduction	815
13.2. One-Sided Search	819
13.2.1 One-sided buyer search: Theory	820
13.2.2 One-sided buyer search: Empirics	821
13.2.3 One-sided seller search: Theory	822
13.2.4 The empirics of one-sided search	822
13.2.4.1 Duration	822
13.2.4.2 Seller motivation	823
13.2.4.3 Seller equity	823
13.2.4.4 Asking price	824
13.2.5 Conclusion	824
13.3. Random Matching	825
13.3.1 Random matching: Basics	826
13.3.2 Intensity choice	828
13.3.3 Entry	829
13.3.4 Turnover and the joint buyer–seller problem	829
13.3.5 Opportunistic matching	831
13.3.6 Seasonality and market thickness	831
13.3.7 Liquidity	832
13.3.8 Behavioral issues	834
13.3.9 Conclusion	834
13.4. Pre-search, Focused Search, and Segmented Search	835
13.4.1 The internet and housing	835
13.4.2 Focused search	837
13.4.3 Segmented search	838
13.5. Directed Search for Housing	839
13.5.1 Asking price as a strategic instrument	840
13.5.2 Asking price as a commitment	840
13.5.3 Asking price as a partial commitment	841
13.5.4 Asking price as a signal	842
13.5.5 Directed search and bargaining	842
13.5.6 Structural approaches to directed search	843

13.5.7	Directed search and booms and busts	844
13.5.8	Housing search and labor markets	845
13.6.	Auctions	845
13.6.1	Theoretical issues	846
13.6.2	Empirics	847
13.6.2.1	<i>Auctions and outcomes</i>	847
13.6.2.2	<i>Bidding wars</i>	848
13.7.	Real Estate Brokers: Fundamentals	850
13.7.1	The activities of real estate brokers	850
13.7.2	Why do real estate brokers exist?	852
13.8.	Competition in the Residential Real Estate Brokerage Industry	855
13.8.1	Nature of competition	855
13.8.1.1	<i>Entry</i>	855
13.8.1.2	<i>Product differentiation</i>	857
13.8.1.3	<i>Price competition</i>	858
13.8.2	The “commission puzzle” and the competitiveness of brokerage	859
13.8.2.1	<i>Theoretical work</i>	859
13.8.2.2	<i>Empirical work</i>	861
13.8.3	Social inefficiency	863
13.8.3.1	<i>Theoretical work</i>	863
13.8.3.2	<i>Empirical work</i>	863
13.9.	Incentive Issues in Real Estate Brokerage	865
13.9.1	Incentive misalignment in real estate brokerage: Theory	866
13.9.2	Empirical work on incentive misalignment	869
13.9.2.1	<i>Broker-owned versus client-owned properties</i>	869
13.9.2.2	<i>Broker-listed versus FSBO properties</i>	870
13.9.2.3	<i>Traditional brokers versus discounted brokers</i>	871
13.9.2.4	<i>Exclusive agency versus exclusive right to sell brokers</i>	872
13.9.2.5	<i>Broker representation</i>	873
13.9.2.6	<i>Full-commission brokers versus split-commission brokers</i>	875
13.9.2.7	<i>Incentive issues in rental markets</i>	876
13.9.2.8	<i>Conclusion</i>	877
13.9.3	Mitigating the incentive problems	877
13.10.	Conclusions	878
	Acknowledgments	879
	References	879

Abstract

This chapter surveys the literature on the microstructure of housing markets. It considers one-sided search, random matching, and directed search models. It also examines the bargaining that takes place once a match has occurred, with the bargaining taking various forms, including two-party negotiations of different types and multiparty housing auctions. The chapter reviews research on real estate brokers as intermediaries as well, focusing on the role of brokers in the matching and bargaining process, the nature of competition and entry in the brokerage industry, and the incentive issues that are present. The chapter also considers the inefficiencies that pervade the brokerage industry and the related policy debates. These are important issues both because of the inherent importance of housing and brokerage and because of the importance of housing to macroeconomic dynamics.

Keywords

Real estate, Housing, Search and matching, Intermediation, Bargaining, Incentives, Cycles

JEL Classification Codes

D82, D83, E32, L85, R21, R31

13.1. INTRODUCTION

There is a compelling case for learning more about housing markets. Housing is typically the largest element of a household's portfolio, so it is central to household finance. Coupled with the illiquidity of housing, this creates a strong empirical relationship between housing markets and labor markets. Housing is also a large part of aggregate capital. Coupled with housing's riskiness, this makes housing an important driver of the business cycle. Housing has played an especially large and unhappy role in the recent global recession. It is clear that households, investors, economic policymakers, and economists themselves all had incomplete understandings of how profoundly housing could affect balance sheets and economic activity more generally.

This chapter will examine one key aspect of housing markets by surveying the growing literature on how housing markets clear. The chapter covers the entire process of buying and selling properties. It discusses models of search, matching, and bargaining.¹ It also discusses models of real estate brokerage. The chapter reviews both theoretical work and empirical work. Not surprisingly, information issues are central.²

The analysis of housing markets must begin with the recognition that housing is a unique good. This means that the analysis of other markets cannot simply be brought to bear on the housing market without modification. This is not a new observation. Various previous surveys of housing have been organized around lists of key housing market features (e.g., [Quigley, 1979](#); [Arnott, 1987](#)). We are interested here in the features of housing as an economic good, the key agents involved in housing markets, and the institutions that govern how housing is transacted.

Three aspects of housing are most important for our purposes. First, houses are heterogeneous. This is true with regard to both the many characteristics of housing units (i.e., square feet of structure and of lot, number of rooms, and age) and the neighborhood amenities and local public goods that are jointly consumed with housing (e.g., access

¹ Most of the literature deals with transacting a fixed stock of houses, so development and redevelopment will not be central themes of the chapter. Even so, some papers consider construction, which will interact with search in a way that has an important impact on housing dynamics.

² There is also a compelling case for learning about commercial real estate. Unfortunately, despite commercial markets involving similar frictions to residential markets, there has been very little research on search, matching, and intermediation in this setting. There is a clear need for more research in this area.

to work and school quality). This heterogeneity means that housing markets can be thin. Second, housing transactions take place under uncertainty. Buyers who are considering moving do not know which houses will suit their tastes until they search. Sellers who are considering moving do not know which buyers might be good matches for the houses that they hope to sell. As a result, buyers and sellers must search for each other. When they meet, neither buyers nor sellers are able to anticipate with certainty the price that the other party in the transaction is willing to pay or accept. Third, there are important market frictions. Search is a costly activity. In addition, the exchange of housing has important transaction costs, including brokerage fees, transaction taxes, and moving costs. In this situation, the housing market clears through price and time (unlike standard competitive models). This liquidity issue will have very important implications for the analysis of housing, as will be seen below.

The housing market is also unique in its institutions and in the economics agents involved in it. One very important aspect of housing is that it is a market dominated by amateurs. Roughly two-thirds of North Americans are homeowners, and these houses are a very important element in a typical household portfolio (see [Tracy and Schneider, 2001](#) for US evidence). A significant share of these houses is transacted with the assistance of real estate agents ([National Association of Realtors, 2005](#)).³ The listing agent is employed by the seller, typically earning a share of the sales prices as a commission. Although the cooperating agent works with the buyer, in a legal sense, the cooperating agent is usually a subagent of the listing agent. It is immediately apparent that there are potential incentive issues here in the use of these intermediaries to facilitate housing transactions. In North America, real estate agents make use of the multiple listing service (MLS) in helping their clients buy and sell houses. The MLS is typically controlled by real estate agent associations (boards) for the use of member agents only. There is controversy about how well this arrangement serves house buyers and sellers. It has been argued that this creates market power, resulting in higher costs of intermediation. On the other hand, it has also been argued that this resource has resulted in sellers and buyers having better information, which might lead to better market outcomes.

This chapter surveys the empirical and theoretical literatures on the microstructure of housing markets. It will consider models that respect the unique features of housing, will assess their ability to explain important facts, and will help researchers avoid certain logical fallacies that follow from analysis that lacks solid microfoundations. [Figure 13.1](#) presents a

³ In this chapter, we will follow the common practice of using the terms “broker” and “agent” interchangeably. In fact, there is a formal legal distinction between brokers and agents. The licensing requirements to be a broker are more strict, typically requiring both additional experience as an agent and additional education.

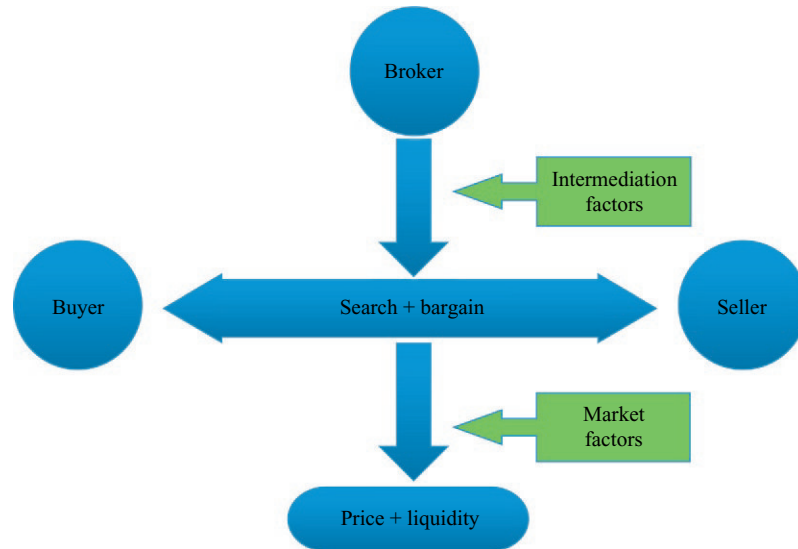


Figure 13.1 Housing market microstructure. *Note: The figure illustrates the key participants in housing transactions (buyers, sellers, and brokers) and the key outcomes (price and liquidity). These interact depending on market factors (listed in Figure 13.2) and intermediation factors (listed in Figure 13.3).*

graphical approach that can be used to understand how the literature has evolved. As with any market, there are buyers and sellers. There may also be a third type of agent, the real estate brokers who intermediate transactions. These agents interact in a number of ways, with the matching between buyers and sellers and the bargaining that ensues being central. The key outcomes are the price at which the house sells and the liquidity of the transaction (if one takes place), typically measured by time-on-market. The research reviewed here considers how these outcomes are affected by a range of market and strategic forces that impact search, matching, and bargaining and also forces that impact intermediation. The former (denoted “Market factors” in the figure) include seller characteristics, house characteristics, market conditions, and a range of strategic choices made by buyers and sellers. Figure 13.2 gives a selective summary of the literature related these factors. Sections 13.2–13.6 discuss this research. The latter (denoted “Intermediation factors” in the figure) include various factors that affect broker performance, such as the ownership of a property, the use of a broker, brokerage representation, the exclusivity and duration of contracts, the nature of compensation structure, and possible incentive alignment issues. Figure 13.3 selectively summarizes research on these factors. Sections 13.7–13.9 discuss this research and also related research on entry, competition, and efficiency in the brokerage industry.

The theoretical literature has moved from simple partial equilibrium models to increasingly rich general equilibrium models with strong microfoundations. For

<u>Categories</u>	<u>Market factors</u>	<u>Selected literature</u>
<i>Seller characteristics</i>	Seller motivation	Albrecht et al. (2007) and Glower et al. (1988)
	Seller equity	Genesove and Mayer (2001)
<i>House characteristics</i>	Idiosyncrasy of the property	Haurin (1988)
<i>Market conditions</i>	Demand shocks	Genesove and Han (2012a)
	Amplification through entry	Wheaton (1990), Novy-Marx (2009), Annenberg and Bayer (2013), Ngai and Sheedy (2014)
	Seasonality	Ngai and Tenreiro (2009), and Salant (1991)
	Market thinness Cyclical variation	Genesove and Han (2012b) Krainer (2001), Diaz and Jerez (2013), and Head et al. (2014)
<i>Strategy</i>	Asking price	Albrecht et al. (2013) and Han and Strange (2014)
	Bargaining	Merlo and Ortalo-Magné (2004), Merlo-Ortalo-Magné - Rust (2013), and Harding et al. (2003)
	Auctions	Ashenfelter and Genesove (1992), Han and Strange (2013)
	Pre-search	Williams (2014) and Piazzesi et al. (2013)
	Advertising	Carrillo (2012)

Figure 13.2 Market factors. *Note: This figure illustrates a range of market factors that the literature has considered. The papers listed are examples only; see the text for a more complete review.*

<u>Intermediation factors</u>	<u>Selected literature</u>
Broker-owned vs. client-owned properties	Rutherford et al. (2005) and Levitt and Syverson (2008a)
Broker-listed vs. FSBO properties	Hendel et al. (2009)
Traditional vs. discounted brokers	Bernheim and Meer (2013) and Levitt and Syverson (2008b)
Exclusive vs. open agency contracts	Rutherford et al. (2001, 2004) and Bar-Isaac and Gavazza (2014)
Broker representation	Miceli (1991), Yavaş and Colwell (1999), Gardiner et al. (2007), and Han and Hong (2014)
Contract duration	Miceli (1989) and Anglin and Arnott (1991)
Full-commission vs. split-commission brokers	Munneke and Yavaş (2001) and Johnson et al. (2008)
Incentive mitigation: competition	Williams (1998) and Fisher and Yavaş (2010)
Incentive mitigation: broker reputation	Shi and Tapia (2014)

Figure 13.3 Intermediation factors. *Note: This figure illustrates a range of intermediation factors that the literature has considered. The papers listed are examples only; see the text for a more complete review.*

example, in one-sided search models, the process that generates the arrival of counter-parties is typically modeled in an *ad hoc* way. In both random matching and directed search models, in contrast, the arrival is an equilibrium outcome, consistent with maximization and learning by the involved parties. Future progress in this area will continue to require attention to microfoundations and equilibrium.

Clear advances to knowledge of both brokerage and matching have come from taking these microfounded theories to data. There have been significant contributions that have adopted modern approaches to identification in reduced-form settings. There have also been significant contributions arising from taking explicit structural approaches and from calibrations. But these approaches have not been the only source of empirical progress in this area. The field has also benefitted from newly available data sources.⁴ For instance, researchers have used new microdata on the offers and counteroffers for a house, on search activity itself, and on nontraditional marketing strategies such as discount brokers. Further progress will therefore rely on both advances in econometric methods and improvements in data.

The remainder of the chapter is organized as follows. [Section 13.2](#) lays out simple one-sided buyer and seller search models. Despite their simplicity, these models explicate some of the key forces at work. The chapter then moves on to consider richer equilibrium models of search and bargaining. [Section 13.3](#) focuses on random matching models, while [Section 13.4](#) covers models where agents strategically focus their search on segments of the market. [Section 13.5](#) considers directed search. In random matching and directed search, bargaining is usually between one buyer and one seller who have met each other. [Section 13.6](#) discusses auctions as a way to transact houses, both formal auctions taking traditional forms and bidding wars that are auctions in fact if not in name. The last three sections in the body of the chapter consider real estate agents as intermediaries, including coverage of fundamentals ([Section 13.7](#)), the nature of competition ([Section 13.8](#)), and incentives ([Section 13.9](#)). [Section 13.10](#) concludes.

13.2. ONE-SIDED SEARCH

We will begin with one-sided models of search. Despite their simplicity, theoretical models of this sort can be quite insightful in explaining how the market factors from [Figures 13.1 and 13.2](#) impact the key transaction outcomes of price and time-on-market. They can also provide a foundation strong enough for careful structural estimation. The most common sort of one-sided model deals with the house seller's problem. There are, however, other sorts of one-sided model, including models of buyer search and models of real estate agent search for buyers and sellers.

⁴ These new sources include CoreLogic, DataQuick, and Internet-based data from sources such as Trulia, Google, Yelp, and Craigslist, to name just a few.

The earliest model of this kind of which we are aware is presented as a part of Simon's (1955) classic analysis of bounded rationality. The body of this paper deals with satisficing as a consequence of, among other things, decision making under uncertainty. There is a brief discussion of the home seller's problem as an instance of this sort of decision making. In the Appendix, there is a formal model of the home seller's problem where offers arise sequentially from a known distribution. The seller must accept or reject these offers, with no recall of previously rejected offers allowed. Simon establishes conditions under which there exists an optimal "acceptance price." Although it is clear that buyers do not simply arrive randomly bearing take-it-or-leave-it offers that expire immediately and irrevocably on their rejection, Simon's analysis of this stopping problem captures some of the most important features of housing transactions: sellers do not know the value of their houses to potential buyers, and they do not know which buyers will visit and when such a visit might take place.

Characterizing this sort of problem as a kind of search, where the acquisition of information is modeled as an economic process, begins with Stigler (1961, 1962). He considers search in goods and labor markets. He does not discuss housing. More modern treatments of search in housing—all one-sided models—are offered in the seminal models of Courant (1978), Stull (1978), Yinger (1981), Haurin (1988), and Salant (1991).

13.2.1 One-sided buyer search: Theory

The first buyer search model in the urban literature is Courant (1978). This model illustrates both the key building blocks of search models and the key motivation for considering search frictions: that such frictions can explain observed outcomes in ways that standard competitive models cannot.

Consider the following simplified version of Courant's model. Buyers search among houses with a cost of c per search. They choose between neighborhoods $j = 1, 2$.⁵ The value to the buyer of a particular house, x_j , is learned only after visiting. The *ex ante* distribution of x is $f(x)$, which is assumed to be equal across neighborhoods. There are two types of buyers: blacks and whites. Sellers are identical except that some will not sell to blacks, with this behavior being unobservable prior to a visit. Let α_j denote the fraction of discriminating sellers in neighborhood j . This is assumed to be known by the searching buyer. Courant abstracts from pricing by assuming that prices arise from a hedonic equilibrium that is independent of race. We will simply assume that houses are identical and prices are fixed.

Search in this model has the standard rule that a buyer should search until obtaining a house above a quality threshold x^* . In the presence of discrimination, search will be limited by neighborhood. For a sufficiently high value of α_j , blacks will not search in neighborhood j . Search frictions, thus, result in segregation. It is worth noting that the

⁵ Courant actually considered an arbitrary number of neighborhoods, J .

segregation only requires that some agents discriminate, not that they all do. In a competitive model, in contrast, discriminatory preferences would be competed away. Courant shows how transaction costs make it unlikely that this discrimination can be arbitrated away. The welfare economics here are thus different than under perfect competition with complete information (although his treatment of pricing here makes it hard to compare). It is worth noting, however, that this model forces statistical discrimination to manifest itself in neighborhood choices rather than in pricing by the assumption of exogenous pricing.

In addition to generating segregation, Courant shows how search and discrimination can interact to produce neighborhood tipping, when a neighborhood's composition changes catastrophically. The most natural way to see this is to suppose that blacks never refuse to sell to other blacks, but a fraction of whites do. Suppose also that some blacks are willing to search even in white neighborhoods. This could be explained by heterogeneity in income within the black population coupled with heterogeneity in neighborhoods. In this case, as a neighborhood becomes more black, the likelihood falls of a buyer wasting time searching the listing of a discriminating seller. This leads to a critical share of black homeowners beyond which all black households would be willing to search in predominantly white neighborhoods, an instance of tipping.

13.2.2 One-sided buyer search: Empirics

There has been relatively little empirical research on buyer search that relates to the one-sided theory discussed above. This seems to be a consequence of data availability. While it is standard for MLS data to report time-on-market for a seller (see below), there is no parallel for a buyer's time-on-market in any standard data set. As a result, empirical research about the buyer-side search activities would require significant effort in data collection and compiling.

Anglin (1997), Elder et al. (1999), and Genesove and Han (2012a) are examples of this sort of research. Based on a survey of buyers, Anglin found, among other things, that information is central to buyer search, whether measured by buyer time-on-market or the number of houses visited. Using cross-sectional data from the National Association of Realtors' (NAR) buyers and sellers surveys, Elder et al. considered the number of houses seen per week. They showed that less-informed but arguably more motivated out-of-town buyers search more intensively. More recently, Genesove and Han aggregated microdata from the NAR buyers and sellers surveys to the metropolitan statistical area (MSA) level, for available years from 1987 to 2008, to form a panel data set of home search activities. They considered buyer time-on-market and the number of home visits, in a random matching two-sided model that also considers sellers. Since buyers are arguably more active than sellers are, empirical research on buyer search is essential for reaching a better understanding of housing markets.

13.2.3 One-sided seller search: Theory

There is a much larger body of work that models the home seller's problem as a process of one-sided seller search. [Stull \(1978\)](#) and [Salant \(1991\)](#) are two particularly influential papers. [Stull \(1978\)](#) considered this sort of trade-off for a rental landlord. [Salant \(1991\)](#) considered the sales process for a homeowner.

Housing search models in this tradition typically have a number of common features, and we will work with a hybrid. There is one house seller facing a sequence of randomly arriving buyers. The seller sets an asking price. This is interpreted as a commitment in the sense that a sale occurs when a randomly arriving buyer is willing to accept. In this setup, there is essentially a take-it-or-leave-it offer of the asking price that the randomly arriving buyer considers. Other bargaining setups are obviously possible, and they will be considered below.

These models all generate some version of the fundamental asking price trade-off: a lower asking price results not only in a lower sales price (by construction) but also in an earlier expected sale. In any search model, markets clear on both price and time. In the case of the housing market, the time dimension captures the liquidity of housing assets and the frictional rate of vacancy of housing markets.

[Salant's \(1991\)](#) model adds some additional dynamic issues to the analysis. He places housing search in a nonstationary setting with a fixed (spring and summer) house sale season. This leads to asking price declining over the season despite his maintaining the usual assumption that there is a constant probability of a buyer arrival with buyers drawing idiosyncratic match values upon their visits.⁶ Overall, there has been relatively little research on housing search in a nonstationary setting. [Merlo et al. \(2013\)](#), discussed below, is a recent exception.

13.2.4 The empirics of one-sided search

13.2.4.1 Duration

There is an extensive literature on the price–liquidity relationship that the theory of seller search motivates. [Cubbin \(1974\)](#) and [Miller \(1978\)](#) are early papers that show an empirical relationship between price and seller time-on-market.

[Haurin \(1988\)](#) is seminal in this line of research. He considers the heterogeneity of housing in a one-sided search setting. He is specifically interested in the atypicality of a particular house. As Haurin notes, some houses have unusual features such as swimming pools in cold climates or unusual ratios of bedrooms to bathrooms. In Haurin's analysis, this affects the variance of the value of a house but not the mean. It is straightforward to see that this leads to a stricter stopping rule (higher price offer to persuade a seller to

⁶ Salant also considered the decision of when to hire an agent, an issue that will be considered later in the chapter.

transact) and a longer marketing time. Haurin makes use of a small data set (219 transactions) from Columbus, Ohio, to assess the empirical validity of these predictions. It has been usual in this literature to work with small data sets from one market. This is because larger cross-sectional or panel data have previously lacked the hedonic characteristics that are needed. This has changed.

Empirically, atypicality is measured as follows. Let β_i denote the hedonic price of the i th attribute in a sample. Let $\bar{x}_i$ denote the average of the i th attribute. The atypicality of a given home is then defined as the sum of the absolute values of the differences in its attributes from the mean weighted by the hedonic prices of the attributes, $\sum \beta_i |\bar{x}_i - x_i|$. This measure of atypicality has become the standard approach in housing research.⁷ Haurin's principal conclusion is that less typical houses do take longer to market, as predicted.

13.2.4.2 Seller motivation

A long list of other papers has considered related liquidity (time-on-market) issues with a one-sided seller search model as the theoretical foundation. Zuehlke (1987) showed that vacant houses exhibit positive duration dependence, with the hazard of sale rising with time-on-market. This result is interpreted as arising from the keen motivation of this group of sellers. Glower et al. (1998) considered seller motivation directly. They employ data on planned moves, and they find that sellers who plan moves sell sooner than those who do not. They also showed that the time to the planned move is associated with seller time-on-market, with sellers planning to move sooner choosing to sell sooner. This result follows naturally from a one-sided model. Springer (1996) also considered motivation, again reaching the conclusion that motivated sellers sell more quickly.

13.2.4.3 Seller equity

Genesove and Mayer (1997) did not write down an explicit search model, but their results can be understood as fitting in this tradition. They considered the impact of homeowner equity on the home sale process. It is straightforward to extend the above seller model to include sellers who vary in their reservation prices. One source of this heterogeneity is that some sellers may have little or even negative equity in their houses. In this case, they will only be willing to sell if they encounter a high-willingness-to-pay buyer. This will lead the sales price of a house to be higher and the time-on-market to be longer when there are equity constraints. They tested these predictions using a unique sample of matched buyers and sellers. The positive relationship between the seller's loan-to-value ratio and duration is shown to be quite robust.

⁷ This measure is frequently used to measure heterogeneity, including the recent papers by Haurin et al. (2010), Bar-Isaac and Gavazza (2014), and Han and Strange (2014).

13.2.4.4 *Asking price*

In the basic seller search model sketched above, the bargaining process is a simple one. Housing is treated as a posted-price market, where buyers arrive and are either unwilling to pay the seller's asking price (in which case, search goes on) or willing (in which case, there is a transaction and the search ends). While this approach is conveniently tractable, it does not capture the bargaining stage of housing transactions. We will deal with this issue extensively below. For now, we will deal only with the line of empirical research that treats asking price as a price posting and looks at the relationship between asking price and the search process.

A long line of research of this sort has dealt with the relationship between the asking price and the outcome variables of sales price and time-on-market. See, for instance, [Miller and Sklarz \(1987\)](#). The idea is that an overpriced house will be less attractive to visitors and so will be on the market for a longer period. It will, however, sell for a greater price. This approach is clearly a reasonable way to think about search in a goods market. If a grocer advertises a high price for apples, fewer customers will visit the store to inspect the apples and decide if they are worth buying. However, a house is different from an apple in many ways, one of them being that houses do not always sell at their posted prices (although a notable fraction do, as observed in [Han and Strange, 2014](#)). This means that the theoretical link between overpricing and outcomes requires some further thought. See also [Pryce \(2011\)](#) for the observation that markets will vary in what the asking price means, implying that measures of overpricing should capture these cross market differences. Having said this, the reduced-form empirical literature clearly establishes that there exists a robust relationship between various measures of overpricing and outcomes. [Yavaş and Yang \(1995\)](#) showed that a higher asking price, controlling for characteristics, leads to a longer time to sale. [Anglin et al. \(2003\)](#) showed that the difference between asking price and a hedonic prediction of sales price is related to longer time to sale. Later in the chapter, we will discuss theoretical models of the interaction between search and bargaining that can help to understand these important results.

13.2.5 Conclusion

Taken individually, each of the one-sided approaches discussed above has plausible microfoundations. Buyers clearly must incur costs to evaluate houses for purchase (or rent, although renting is typically not considered). Sellers incur costs too: in the extreme, staging; in a less extreme case, preparing a house for viewing by cleaning, tidying, buying flowers, and absenting oneself. And finding an agent has its own set of uncertainties and search costs, which could be understood using the same sorts of model. Search models seem a very natural way to capture these situations. Versions of the key results from these one-sided models will reappear throughout the chapter.

However, one-sided models by construction take the process that generates the arrival of counterparties as exogenous. In [Merlo et al. \(2013\)](#), the setting is in the United Kingdom, where the process of search does indeed seem to be well described as one of one-sided seller search. Their structural analysis, discussed in greater detail below, solves for optimal search behavior for sellers based on an empirically derived process of offer arrival. This insightful approach requires data on offers, which is difficult to obtain. In other one-sided models, the counterparty arrival process is arbitrary, which is not very satisfying. [Section 13.3](#) will begin to consider general equilibrium approaches where both buyers and sellers optimize their search.

13.3. RANDOM MATCHING

This section considers random matching models of housing search. It also considers the bargaining that takes place after buyers and sellers meet, since bargaining is an inseparable part of the search and matching process. As in [Figures 13.1 and 13.2](#), it will consider a range of market factors that influence housing transactions, with additional factors considered later in the chapter.

The models discussed in this section are adapted from labor economics, where search models have proven to be remarkably useful. The work of Diamond, Mortensen, and Pissarides is seminal. See, for instance, [Diamond \(1982\)](#), [Mortensen \(1982\)](#), [Mortensen and Pissarides \(1994\)](#), and [Pissarides \(1985\)](#), among others. This vast literature has been surveyed many times, including [Pissarides \(2002\)](#), [Rogerson et al. \(2005\)](#), [Shi \(2008\)](#), and more recently [Rogerson and Shimer \(2011\)](#), and we will not attempt such a survey here. We will focus instead on how models in this tradition have been applied to housing economics.

The labor literature has shown that search models can help to understand a number of fundamental issues about labor economics. As noted in [Rogerson et al. \(2005\)](#), these include the duration of unemployment spells, the simultaneous existence of unemployment and job vacancies, the determination of unemployment and turnover, and the interaction of wage and turnover in equilibrium. There are parallels for all of these issues in housing. These include time-on-market, housing vacancies and turnover, and the dynamics of price, construction, and liquidity. As will be seen below, random matching models of housing markets illuminate all of these issues.

Random matching models also help to understand some of the key stylized facts of housing dynamics and thus contribute to the understanding of the great boom and bust that took place in the 2000s. Different authors have focused on different facts. As a taxonomy, we see three broad types of fact that have been considered. First, there is a tendency toward cycles in housing markets in the sense that there is short-run positive serial correlation of price but long-run mean reversion (see [Case and Shiller, 1988](#)). Construction also shows persistence ([Glaeser and Nathanson, 2014](#)). Second, shocks to

fundamentals result in amplified effects on housing prices and quantities, an instance of “excess volatility” (Shiller, 1982; Glaeser et al., 2014). Third, there are regularities in how price, sales, and various dimensions of liquidity such as time-on-market move together across time. Price and sales exhibit positive correlation, while price and seller time-on-market show negative correlation (Stein, 1995; Krainer, 2001; Glaeser et al., 2014). See Glaeser and Nathanson (2014) for an extensive discussion of how these facts relate to the phenomenon of housing bubbles.

The parallels between housing search and other sorts of search are sometimes not exact, however. There are issues unique to housing that must be addressed. One of these is that buyers of houses are also sellers. This means that the search frictions in housing markets are different and more serious than in labor markets. This has been shown to lead to increases in volatility. It also means that the market clears on price and both buyer liquidity and seller liquidity. This has implications for empirical work. Another key difference is that there is bargaining over price. While there is also bargaining over wage, this is not a feature of most labor search models. That the asking price is not exactly a posted price will matter, both theoretically and empirically. Furthermore, the role of intermediaries in housing search is absolutely central, while labor market intermediation is secondary. We will deal with all of these in this section and later in the chapter.

13.3.1 Random matching: Basics

The following is a simple example containing the key elements of a random match model of housing. Although it is derived from Genesove and Han (2012a), there are numerous other instances of this kind of model in the housing literature.⁸ Let n_b and n_s give the numbers of buyers and sellers in a market. Both are assumed to be risk neutral. Let $\theta = n_b/n_s$ denote market tightness. Let $m(n_b, n_s)$ denote the rate of meetings or contacts between counterparties. We suppose that $m(-)$ exhibits constant returns, as is conventional. Not all contracts will lead to matches, since some matches are not productive enough. Whether a contact leads to a successful transaction depends on the total surplus that the contact generates.⁹

In this setup, the probability of a given seller being contacted by a random buyer can be expressed as

$$q_s(\theta) = m(n_b, n_s)/n_s = m(\theta, 1). \quad (13.1)$$

Similarly, the probability of a given buyer being contacted by a random seller can be expressed as

⁸ See, for instance, Wheaton (1990), Krainer (2001), Albrecht et al. (2007), and Novy-Marx (2009).

⁹ See Pissarides (2000) for a discussion of the concepts of meeting and matching functions. The former refers to contacts between agents, while the latter refers to successful contacts that result in ongoing matches.

$$q_b(\theta) = m(n_b, n_s)/n_b = m(1, 1/\theta). \quad (13.2)$$

We then have $q_s(\theta) = \theta q_b(\theta)$. Assuming that $m(-)$ is increasing in both arguments gives $q'_s(\theta) > 0$ and $q'_b(\theta) < 0$. When a contact is made, the idiosyncratic match utility of a particular house for a particular buyer, x_{ij} , is realized. *Ex ante*, it is a draw from $g(x)$. This distribution is known to both buyers and sellers. Define $\gamma = V_b + V_s$ as the sum of the reservation values of the buyer and the seller. We will treat V_b as exogenous, which means that there are many markets from which a buyer might choose. Then, the surplus from a match equals $x - \gamma$. The probability of a transaction conditional on a meeting equals $G(\gamma) = \text{prob}(x - \gamma \geq 0)$. The expected surplus conditional on a transaction equals $E(x|x - \gamma \geq 0) - \gamma$. When a meeting occurs, the seller and buyer engage in Nash bargaining over the potential surplus from the transaction. Let β represent seller bargaining power and $(1 - \beta)$ represent buyer bargaining power. In this case, given a transaction, the expected price will be

$$p = V_s + \beta^*[E(x|x - \gamma \geq 0) - \gamma]. \quad (13.3)$$

Finally, let the cost of a search be c_b and c_s for buyers and sellers, respectively, and let r be the interest rate.

The endogenous variables are γ and θ . In equilibrium, the asset equations governing search are

$$rV_s = q_s(\theta)G(\gamma)^*\beta^*[E(x|x - \gamma \geq 0) - \gamma] - c_s, \quad (13.4)$$

$$rV_b = q_b(\theta)G(\gamma)^*(1 - \beta)^*[E(x|x - \gamma \geq 0) - \gamma] - c_b. \quad (13.5)$$

Equation (13.4) requires that the opportunity cost of a seller continuing to search for another period must equal the net benefit of seller search. The latter equals the probability of the seller matching multiplied by the probability of a meeting resulting in a transaction multiplied by the price that the seller receives, minus the search cost. Equation (13.5) is similarly a requirement that the opportunity cost of buyer search equals the expected net benefit accruing to buyers from continuing to search. It is worth observing that without the convenient assumption of fixed buyer utility, there would be a third endogenous variable, buyer utility V_b . This would be determined by an entry condition for buyers. Some of the models discussed below take this approach.

This basic random matching model can generate some of the stylized facts discussed above. It gives amplification in the sense that the short-run adjustment to a shock can generate overshooting. In addition, this basic random matching model can generate a positive price–volume correlation. The model requires modification in order to generate the persistence mentioned at the beginning of the section. We will return to this later.

The model also makes predictions about how housing market liquidity—measured by buyer time-on-market, seller time-on-market, and the number of visits made by a

buyer—would respond to shocks. A positive demand shock that increases the expected surplus from a transaction increases the buyer to seller ratio (market tightness), which further increases the seller contact hazard $q_s(\theta)$ but decreases the buyer contact hazard $q_b(\theta)$. Such demand shock also increases the probability of a transaction conditional on a contact, making each home visit more productive. As a result, a positive demand shock would decrease both the seller time-on-market and the number of home visits that a buyer makes, but its net effect on buyer time-on-market is ambiguous.

Using National Association of Realtors' buyers and sellers surveys from 1987 to 2008, [Genesove and Han \(2012a\)](#) assess these predictions. They find that an increase in income or in population has substantive and negative effects on seller time-on-market and the number of visits made by a buyer. However, these effects are much smaller and less significant in the long run. These results are consistent with a random matching model where sellers react to demand with a lag and interest rate is effectively negligible. Their findings also imply that the elasticity of the hazard that any given seller will be contacted by a buyer with respect to the buyer–seller ratio is 0.84, assuming that a constant returns to scale matching function. Thus, a doubling of the buyer–seller ratio would increase the likelihood of any given seller being contacted by a buyer by 79%.¹⁰ Such a large seller contact hazard elasticity is consistent with the seller listing institutions in North American real estate markets.

13.3.2 Intensity choice

The basic random matching model has been extended in numerous ways (recall [Figures 13.1 and 13.2](#)). We will consider several of these, beginning with the choice of search intensity.

Suppose that buyers and sellers choose their intensity of search and that this will impact the likelihood of a match taking place. The match probability for a given seller now depends on the market tightness, own effort, the effort of other sellers, and the effort of buyers. Own effort and the effort of buyers increase the probability of a match, while the effort of other sellers decreases the probability. The setup for buyers is parallel. In this situation, the comparative static results discussed above for market liquidity are preserved under certain restrictions on the modified matching function. Loosely, it suffices that the marginal cost of effort rises sufficiently quickly and that seller effort is sufficiently dissipative.¹¹

¹⁰ It is worth noting that [Head et al. \(2014\)](#) arrive at a quite similar estimate using very different data and methods. We discuss this paper below.

¹¹ Seller effort is dissipative in that increasing seller effort would only steal buyers away from other sellers without much effect on improving the overall matching rate. See [Genesove and Han \(2012a\)](#).

13.3.3 Entry

The basic random matching model above treats buyer utility as exogenous. [Novy-Marx \(2009\)](#) supposed instead that both buyer entry and seller entry depend on the payoffs of, respectively, buying and selling a house. He does this by supposing exogenous functions giving the number (literally, measure) of house buyers and sellers as functions of price. These functions capture outside alternatives. For instance, housing construction is almost certain to grow when the price rises. Furthermore, a buyer would presumably be more willing to move to another market if his/her house would sell for more. Of course, if the other market was positively correlated, this incentive would be weaker (see [Han, 2010](#), for a discussion of correlations and the management of risks in housing transactions).

This setup then generates the important result that these natural entry processes lead to an amplification of shocks. In Novy-Marx's usage, there are "hot and cold markets." The amplification works as follows. Suppose that there is a shock to demand, perhaps from a change in mortgage standards allowing more households to qualify for mortgages. This increase in the rate of buyer entry leads to a shorter time-on-market. This, then, further increases the ratio of buyers to sellers, amplifying the effect of the initial increase in buyer entry. A similar amplification story would apply in cold markets. This result shows how the search structure of housing markets contributes to their often-observed cyclical nature.

In all of these models, buyers and sellers are different agents. In fact, as discussed above, one of the most important unique features of housing is that buyers are sellers. As long as the inflow of such "dual" agents had no effect on the net inflow of buyers less sellers, ignoring them would not be consequential to the steady-state analysis. However, it is possible that for such "dual" agents, their subsequent actions as buyers or sellers are strongly interrelated, which would affect the resulting house price and market liquidity. We will now turn to the seminal model of this sort.

13.3.4 Turnover and the joint buyer-seller problem

[Wheaton \(1990\)](#) presented a model of what might be called "in-home search." The model is parallel to models of labor search where workers consider possible alternate positions while they are on the job. The crucial innovation in this model is its recognition that homebuyers are frequently home sellers. It is, thus, a model of turnover. That buyers are sellers and that these roles are linked is an important difference between housing search and labor search.

Wheaton's model considers this link as financial. There are two types of houses and households. Households suffer a utility penalty if they are occupying a house that is not well matched to their type. Random shocks move households into a poorly matched state, requiring them to buy a new house. After they do this, they occupy two houses, one of them vacant, until they sell it. While this approach ignores some of the issues

involved in changing houses, it captures the crucial fact that a given agent operates as both a seller and a buyer. The other elements of the model are as in the above random matching framework: there is random contact between agents that depends on effort. One important simplification is that some of the analysis supposes that the matching probability depends only on effort and not on vacancy.

The model generates some of the key stylized facts of the housing market: the existence of structural vacancy in steady state and the possibility of amplification, where a small shock to fundamentals produces large movements in price. In this case, because buyers are also sellers, the price needs to rise substantially to equilibrate in response to a demand shock. This channel through which search can affect the amplitude of cycles depends critically on the fact that in housing, turnover means something different than in other search markets.

Turnover is also central to a recent paper by [Anenberg and Bayer \(2013\)](#). They considered a more specific sort of turnover than Wheaton, situations where households buy and sell one house within a market. They showed that this internal turnover is highly volatile and, in fact, the primary element of housing volatility. They then developed a search model to incorporate this. For internal moves, the cost of holding two houses varies endogenously over the cycle. They estimate the model using Los Angeles data on moves, prices, and marketing times. The empirical analysis shows that this turnover issue amplifies the real estate cycle.

[Head et al. \(2014\)](#) worked with a calibrated search model that examines the dynamics of house prices, sales, and construction and the entry of buyers in response to city-specific income shocks. This allows them to assess the ability of such a model to quantitatively match the key stylized facts for the US housing markets. In their model, the entry of new buyers and the construction of new houses are endogenously determined. Every homebuyer eventually becomes a seller through either a failed match with the city she lives in or a failed match with the house she currently owns. This results in the movement of the owner to temporary renting, searching for a new house, and the owned house being put on the market. With these new features, Head et al. showed that a random matching model can generate short-term persistence in price growth and construction even in the absence of persistent income growth. Quantitatively, their calibrated model accounts for over 80% of house price volatility driven by city-level income shocks and nearly half of observed serial correlation in-house price growth.

[Ngai and Sheedy \(2014\)](#) employed a model that endogenizes the moving decision. In the other models considered here, such as [Head et al. \(2014\)](#), matches between homeowner and house become unproductive according to a random process. By assumption, when a match becomes unproductive, the household must move. Ngai and Sheedy endogenized moving by supposing that a household chooses to move according to the stochastic quality of the match, which is the outcome of series of smaller shocks rather than a discrete and total loss of match quality. The critical match quality at which a household

moves is solved for as an equilibrium level. This leads to amplification, since a shock leading some households to move will encourage further moves due to thick market effects. It also leads to what they label as “cleansing,” since the remaining matches are those of higher quality. This means that there is overshooting as the economy adjusts to a new steady state.

This line of literature is quite important. Housing markets are notoriously cyclical. The papers discussed above help to understand housing cycles in several related ways. They establish that small changes in fundamentals can lead to large changes in outcomes, which helps to explain the volatility of real estate. They also help to explain some of the key cyclical properties of housing markets. These include the persistence of house price changes in the short run and the comovements across the cycle of price, volume, liquidity, and vacancy. Given the profound impacts of the recent boom and bust cycle, there is clearly a strong case for further research in this area.

13.3.5 Opportunistic matching

[Albrecht et al. \(2007\)](#) presented a model that captures the increasing desperation that seems to characterize housing search. Their model starts with the standard elements of matching models. They captured desperation in a creative and elegant fashion by assuming that agents begin as relaxed, with high flow values from being in an unmatched state and continuing to search, and then become desperate according to a Poisson process. They showed that there are various forms that the equilibrium can take, including indiscriminate matching (where all matches are consummated), opportunistic matching, where desperate agents match with everyone but relaxed ones wait in the hope of encountering a desperate counterparty. There is also a type of equilibrium where only the desperate match.

The model fits with empirical observation in various ways, including the relationship between time-on-market and price. The two-state structure of the model (relaxed or desperate) allows for very sharp characterizations of equilibrium strategies. Of course, this kind of exogenous change in payoffs is not the only transition that might take place for a buyer and seller during a housing search. There could be changes in market circumstances or learning. And the relaxed/desperate states could be interpreted as capturing buyer and seller rationality or irrationality.

13.3.6 Seasonality and market thickness

The above analysis has considered constant returns matching functions. It is possible that matching functions exhibit increasing returns, in which case there would be thick market effects (see [Petrongolo and Pissarides, 2001](#)). There are several potential sources of variations in thickness across markets. One is the classic urban economic phenomenon of agglomeration: larger cities have thicker markets. Another kind of variation occurs across

the business cycle, with volume varying cyclically. Yet another is found in the well-known seasonality of real estate markets. This was captured in [Salant's \(1991\)](#) one-sided, nonstationary model of the home seller problem.

More recently, [Ngai and Tenreiro \(2014\)](#) carried out a search and matching analysis of this issue. The paper is motivated by the preference of buyers and sellers to exchange houses during the “hot” second and third quarters of the year. Using UK data, the seasonalities are documented and shown to be parallel to similar patterns documented in other markets. This is used to motivate a model where households have exogenous changes in move preference by season due to the school calendar and other factors. In this search and matching model, these exogenous preferences generate thick market effects, thus amplifying the effects of shocks. The model is calibrated, and the calibrated model accounts for seasonal fluctuations in both the United Kingdom and the United States.¹²

13.3.7 Liquidity

A common feature of every model reviewed so far (and many of them that will follow) is that housing markets clear on both liquidity and price. This has clear implications for the construction and interpretation of house price indices. In particular, the market frictions captured by search predict that there is a relationship between price and time-on-market for both buyers and sellers.

[Krainer \(2001\)](#) is a classic model in this spirit. He presents a random matching model that predicts cyclical liquidity of owner-occupied houses.¹³ In the basic version of the model, sellers have a high opportunity cost of failing to sell during a given period, and this encourages more rapid sales. A buyer's opportunity cost of mismatch—continuing to search—is assumed to equal the loss of housing consumption. Since this is greater by assumption in a boom, buyers also transact more rapidly in a boom. Interestingly, Krainer showed that with a frictionless rental market, these liquidity effects are not present, since it is possible to benefit from the boom without a purchase. Since the simultaneous existence of rental and purchase markets is an important feature of housing markets, this is a valuable result.

The relationship between liquidity and price raises immediate and fundamental questions. Standard assessment practice is that a house's assessed value should equal the amount that the house would sell for in an arms' length transaction taking a reasonable amount of time. But what is reasonable? And if marketing time depends on the state of the market, then what can one say about the performance of price indices? Furthermore, it has long been recognized that house price can provide information on the value of

¹² [Harding et al. \(2003\)](#) considered estimate bargaining power using American Housing Survey data. Among the interesting results of this paper is that bargaining power is seasonal for families with children, with the pattern consistent with intuition.

¹³ See also [Krainer and LeRoy \(2002\)](#).

unmarketed amenities since the amenities are jointly purchased with the house. What is the implication of search and illiquidity for the interpretation of this information?

Several papers have considered this issue. [Kim \(1992\)](#) showed how search truncates the set of possible transactions and introduced biases into estimates of hedonic coefficients. He proposed a maximum likelihood approach accounting for the truncation to address the bias. He implemented the model on American Housing Survey rental data from Sacramento. One result that is particularly relevant to search is that new residents, who presumably face higher search costs, have higher reservation prices.

[Fisher et al. \(2003\)](#) analyzed commercial real estate indices in a similar context. The core of the paper is a search model of commercial real estate. Buyers and sellers both have reservation prices, and transactions require a match where the buyer's reservation price exceeds the seller's. It is well known that volumes (the inverse of liquidity) vary procyclically. This can be captured in the Fisher et al. model by supposing that shocks have different impacts on buyer and seller reservation prices. In this case, Fisher et al. showed that indices need to be adjusted for liquidity in order to present an accurate picture of the state of the market.¹⁴

They implement an approach to controlling for liquidity that follows [Heckman's \(1979\)](#) sample selection methods. Identification requires variables that impact seller reservation prices but not buyer reservation prices and vice versa. The liquidity-controlled indices show higher appreciation in booms and greater depreciation in busts. Liquidity correction, thus, shows greater volatility than without correction. They find rather large differences, with the liquidity-corrected index having compound appreciation rate of 0.52% over their sample period, while the uncorrected rate is 0.76%. Of course, commercial real estate is a private market where properties are quite heterogeneous, so liquidity is likely to be especially important.

[Goetzmann and Peng \(2006\)](#) presented a related analysis of a residential market that also shows the necessity of a liquidity adjustment. They presented a procedure that is distinct from Heckman's approach, relying instead on the specifics of the error structure in the models of reservation price and sales price. They calculate liquidity-corrected indices using a repeat sales approach for the Los Angeles MSA. As with other approaches, they again found greater volatility in the corrected series, with higher increases in booms and larger decreases in busts. The correction, while statistically significant, is considerably smaller than in the [Fisher et al. \(2003\)](#) analysis of commercial real estate.

The analysis in these papers shows that considering liquidity is required in order to understand how the housing market evolves. Two recent related papers have made

¹⁴ This paper is rare in focusing on search in a commercial context. Commercial markets are clearly thin, arguably more so than residential markets. Data, however, are scarce, and this seems to have led to an absence of research in this area.

important progress in this regard. Carrillo (2013) developed an index of the “heat” of a housing market that comes from an analysis incorporating search and liquidity. Carrillo et al. (2014) showed that this index can predict future housing movements, a result related to the previously noted results on price momentum and the comovements in price and liquidity across the business cycles.

The analysis also suggests the welfare calculations based on hedonic prices are telling only part of the story, since liquidity is ignored. To measure the value of an attribute requires a correction for illiquidity that is comparable to the corrections discussed above for house price indices. Zahirovic-Herbert and Turnbull (2008) carried out such an exercise for school pricing. There is clearly room for further research of this sort.

13.3.8 Behavioral issues

The literature reviewed thus far has taken the standard economic approach of supposing that all agents are rational. The presence of search frictions does, of course, mean that this rationality is different from the rational behavior that one would see in models with perfect information. Even so, all agents are maximizing expected utility in their search, purchase, and sale decisions based on information that is updated in a way that respects the equilibrium strategies of other agents.

There is a small but growing literature that has considered departures from rationality. Shiller (1999) provided a comprehensive survey of various behavioral heuristics that agents might follow that are inconsistent with standard approaches to rationality. Genesove and Mayer (2001) presented data from Boston’s housing market consistent with the presence of loss aversion, one particular behavioral heuristic.

Several recent papers have considered models that add a behavioral component to a search model. Piazzesi and Schneider (2009) showed how a small number of irrationally optimistic agents can generate positive momentum in housing markets. Burnside et al. (2011) calibrated a model where agents’ contacts with each other result in changes in each other’s beliefs and the beliefs in turn influence housing transactions. The model generates a situation where booms are sometimes but not always followed by busts, consistent with empirical observation. Peterson (2012) presented a model of search where agents incorrectly believe that the efficient market theory holds. The model is shown to replicate housing dynamics. See Glaeser and Nathanson (2014) for further discussion of behavioral models of bubbles.

13.3.9 Conclusion

The main conclusion of random matching models of housing markets is that search frictions fundamentally impact how the housing market clears. These frictions can explain the simultaneous existence of vacancies and households actively seeking to move. They can also explain the observed phenomenon of a positive correlation between

price and volume, something that would not necessarily be predicted in a simple competitive model of housing markets. They can also at least in part explain the volatility of housing markets. Finally, they call into question frictionless bid-rent models of housing and land markets. The next section will extend the analysis of search and matching by considering models with various kinds of search direction.

13.4. PRE-SEARCH, FOCUSED SEARCH, AND SEGMENTED SEARCH

It is not literally true that a buyer is equally likely to search all of the houses in a market. It is just as unlikely that a seller will encounter visitors who are random draws from an entire market. Instead, there is an initial round of “pre-search” based on information from advertisements that allows buyers to inspect only a subset of a market’s houses (which can be conceived of as searching within a submarket). This section will consider models of this sort where pre-search activity gives richer microfoundations to the search models. Since the key issue here is how buyers obtain and make use of information in deciding on a search strategy, it is natural to begin by considering the most important new source of information, the Internet.

13.4.1 The internet and housing

The Internet has had a revolutionary impact on society, and these impacts have been felt in housing markets. While it is not possible to buy a house online in the same way that one buys a book, the Internet can certainly facilitate shopping. For instance, one can carry out preliminary investigations online, learning about house and neighborhood characteristics. While this is not costless since it takes time, and while the information gleaned from this kind of pre-search is not exactly the same as what a traditional visit would provide, it is obvious that the search process is impacted by this possibility.¹⁵

Several approaches have been taken to modeling the impact of the Internet. It provides lower cost search in [Ford et al. \(2005\)](#). It provides specific information in [Carrillo \(2008\)](#). It shifts the distribution of match values in [Genesove and Han \(2012a\)](#), [Han and Strange \(2014\)](#), and [Williams \(2014\)](#). All of these approaches can be thought of as modeling an initial stage of the search process, or pre-search.

[Ford et al. \(2005\)](#) documented the empirical relationship between listing a property on the Internet in addition to the standard MLS listing and sales price and seller time-on-market. The empirical analysis is motivated by search theory. The key assumption in the model is that the marginal cost of an additional search is lower for Internet-listed

¹⁵ Of course, pre-search predated the Internet, with paper listings providing information on which visit decisions could be based. See [Anglin \(1997\)](#) for a useful discussion of newspaper advertisements and buyer search.

properties. The model has increasing marginal cost of search of both the online and conventional types, with the recall of previous listings a possibility. The model predicts that buyers will pay more for houses when they search on the Internet, a consequence of the improved match allowed by the superior learning technology. It also predicts that buyers will search longer on the Internet, again a consequence of the greater net benefit of such searches.

The model's predictions are taken to data on house listings. The key empirical exercise is the estimation of seller time-on-market and sales price in a simultaneous system. Since the choice of listing is endogenous, a selection model is estimated. The paper does not explain what exclusion restrictions would give identification, although for inverse Mills' ratio term, it is common to argue that identification comes from the nonlinear functional form of the selection model. The paper finds that properties listed on the Internet sell for more and take longer to do so.

Carrillo (2008) looked at the related issue of how the amount of visual information is related to market outcomes. Carrillo's analysis is motivated by Jovanovic's (1982) model of information unraveling. The idea is that there are good houses and bad ones, with only houses above a critical quality threshold choosing to provide an extensive virtual tour. His analysis yields a positive prediction of the relationship between sales price and visual information. It yields an uncertain prediction for seller time-on-market; more information might lead to more matches for some houses and to fewer for others. In his empirical analysis, Carrillo addresses the identification issue by adopting an instrumental variable approach to the property-level choice of how much information to provide. He employs two instruments, visual information in listings at the agent's firm and the visual information at nearby houses. His 2SLS results show a strong relationship between the provision of visual information and price, with a virtual tour adding 2% to price and 10 additional pictures adding 1.7%. Regarding time-on-market, which has a theoretically ambiguous relationship to visual information, the empirical relationship is strong and negative. Houses with more information sell more quickly, 20% more quickly in the case of virtual tours.

Genesove and Han (2012a) also had results on the Internet, although this is not their primary focus. This paper jointly considers buyer and seller liquidity in the context of a random matching model. The paper examines search over a period when the share of homes that buyers report as having found on the Internet rises from zero to roughly one-third. It shows in an Appendix how the Internet might impact search and matching where the use of Internet is endogenously induced by demand. The OLS empirical analysis shows that if all buyers find homes via the Internet, buyer time-on-market would be 24% greater and the number of home visited would be 30% greater than if none do. Seller time-on-market is insignificant. That buyers search more intensively when they have complementary information sources available is consistent with the search and matching model.

Finally, [Han and Strange \(2014\)](#) considered the impact of the Internet on the likelihood of a house buyer or seller participating in a bidding war with an above-list sales price. The paper shows that buyers who use the Internet are more likely to have purchased their houses through above-list bidding wars than other buyers. This is consistent with the Internet being an improvement in matching technology where matching improvements increase transaction volumes ([Pissarides, 2000](#)) and high transaction volumes are positively associated with bidding wars.

All of the papers discussed here have taken reduced-form approaches. The principal issue to address for future research of this sort is to find a way to achieve quasi-experimental variation in the use of the Internet to buy or sell a house. The difficulty with existing research in this area is that houses sold using the Internet or households searching through the Internet are different in unobserved ways from those houses and households who transact without the Internet. These unobserved characteristics could enter as errors in pricing equations or other models of interest, resulting in biased estimates.

The next section will consider a rigorous theory of how pre-search activities impact real estate markets.

13.4.2 Focused search

[Williams \(2014\)](#) develops a model of “focused search” that moves beyond a purely random search process. The key assumption in his focused-search model is that the pre-search results in a truncation of the distribution of idiosyncratic match values with the new truncated distribution being well approximated by a power law distribution.¹⁶ The use of a power law approximation has been shown to have strong microfoundations in other applications. The approach here delivers sharp results. In particular, with the additional assumption of an isoelastic effort cost, the model generates explicit solutions for key variables in partial equilibrium and makes the analysis of the full steady-state equilibrium considerably more transparent.

The model is used to capture several important real-world aspects of real estate markets. For instance, it distinguishes in a serious way between the existing home market and the market for new construction. There are important empirical differences between these markets. Existing homes involve negotiation over price but do not allow an opportunity to customize. New houses allow customization, but typically do not allow negotiation. The paper characterizes equilibrium when sellers can allocate their search effort across multiple submarkets. The model and accompanying calibration fit some key

¹⁶ See also [Genesove and Han \(2012b\)](#) for a related analysis of prescreening approach. They considered three families of extreme value distributions as well as the generalized Pareto distributions for the new distribution of the idiosyncratic matching value generated by a truncation of the original distribution of matching values.

stylized facts. Existing houses sell for more than comparable new houses. Buyers search less intensively among new houses but buy more frequently. The distinction here between new and old houses and the different opportunities they offer for customization is a feature of housing search that is not present, at least to a significant degree, in other sorts of search markets.¹⁷ This is another example of how the analysis of housing search must respect the particular nature of housing markets.

13.4.3 Segmented search

The issue of market segments is explicitly addressed in [Piazzesi et al. \(2013\)](#). The crucial innovations in their housing search model are the presence of multiple market segments and heterogeneous clienteles.¹⁸ Their approach makes use of search alerts provided to prospective homebuyers by [trulia.com](#).¹⁹ These alerts inform the buyers when houses meeting particular criteria are listed. They can, therefore, capture buyer tastes. There is a clear pattern of geographic segmentation in the raw data. The geographic segmentation is evidenced in three related ways. Buyers search zip codes that are not too far from each other (the maximum distance between the centroids of all zip codes searched has a maximum of 9.8 miles for a buyer who searches at least two zip codes). Buyers also search zip codes that tend to be contiguous and also that tend to satisfy circularity (searching within a distance of a critical location such as a workplace). There is also price segmentation and, to a lesser extent, segmentation by a house's number of bathrooms, which seems to capture a range of quality and quantity characteristics about which buyers have preferences.

Using this novel and rich data set, Piazzesi et al. found substantial variation not only for market outcomes across segments but also for clienteles within and across segments. In addition, inventory and search activity are inversely related across cities, but positively related within most cities. They then calibrated a version of the Diamond–Mortensen–Pissarides random matching model with multiple segments. In equilibrium, the cross-sectional distribution of turnover, inventory, price, and search activity is related to the distribution of preferences, matching technology, and moving shocks. In particular, the interaction of heterogeneous clients is shown to be quantitatively important for understanding housing market activity, measured by variables such as turnover, inventory, and seller time-on-market. The heterogeneity in these measures across market segments also means that the liquidity discount varies across market segments. In the

¹⁷ The used-car market seems to have somewhat similar features. The labor market does not.

¹⁸ [Guasch and Marshall \(1985\)](#) provided an early analysis of vacancies and marketing time across a segmented rental market.

¹⁹ See [Chauvet et al. \(2014\)](#) for another creative use of search query data, the measurement of market movements. The use of online information in empirical research on housing has exploded in recent years. Such data appear to have the potential to at least partly resolve the greatest obstacle to research in this area, the absence of data.

calibrations, this variation is substantial, with more stable segments with fewer turnovers and more popular segments with low inventories exhibiting a much lower discount than other segments.²⁰ See also Landvoigt et al. (2013) for a related paper that presents an assignment model of continuous segmentation in San Diego's housing market. The results are consistent with the relaxation of credit contributing to San Diego's housing boom, especially at the low end of the housing market.

Liu et al. (2014) is another paper that has considered market segmentation. It takes a dynamic approach and defines segments by house size. The empirical work uses Phoenix data. The key empirical finding is that all size segments moved together during the boom, but during the bust, there is a strict hierarchy of declines, with smaller properties declining more quickly.

This section has considered how and where buyers allocate their search effort and how this affects housing market activity and transaction outcomes. The next section will consider the strategic actions of sellers.

13.5. DIRECTED SEARCH FOR HOUSING

With directed search, an agent uses price to influence the matching process. The previous section discussed models where agents choose how intensively to search and where to allocate their search efforts. With directed search, the seller commits to a price, and this plays an important role in the agents' search choices and the matching that ensues.

Peters (1984, 1991) presented seminal directed search models for goods markets. The former shows how search frictions and capacity constraints can resolve the well-known discontinuity in payoffs in Bertrand pricing games. The latter compares *ex ante* price commitments with negotiation in a search and matching setting. It shows that a seller can profit from setting an *ex ante* price when all other sellers are choosing not to do so, a result suggesting the instability of failing to post prices.

Most of the directed search literature is in the context of labor or commodity search. See Rogerson et al. (2005), Rogerson and Shimer (2011), and Shi (2008) for recent surveys. The issues that are most important for our purposes are as follows. First, the posted price impacts search. Second, the equilibrium posted price under some circumstances gives efficiency in the sense of producing a ratio of buyers to sellers in steady state consistent with the Hosios (1990) condition on market efficiency. This efficiency was absent in random matching models without direction, where the Hosios condition prescribes a buyer–seller ratio that will be obtained only accidentally. One version of this result is

²⁰ It is worth pointing out that the working paper of Levitt and Syverson's (2008a) agents report results on the relationship between property characteristics reported in advertisements and sales price and marketing time. This can be interpreted as descriptive evidence on segmented search that is complementary to Piazzesi et al.

found in [Moen \(1997\)](#) and [Shi \(2001\)](#). Third, unemployment and vacancies are simultaneously present in equilibrium. There is very little recognition in this literature of the unique characteristics of housing.

Should we expect these results to extend to housing? Our answer to this question is mostly, but not entirely. Search is an appropriate model of trading frictions, and housing vacancies and more generally illiquidity can be understood by modeling frictions through search. However, housing has important institutional differences, including the use of intermediaries (much more common with housing), that the same agents are both buyers and sellers on the market and the complicated role of asking price. So one can make use of labor search to understand housing markets, but modification is required. Put the other way, the analysis of housing search has the potential to add to the general literature on search.

We now turn to models that deal with the specifics of housing.

13.5.1 Asking price as a strategic instrument

There are two ways that asking price might impact the sales price of a house in a search setting. First, once there is a meeting, asking price may impact the bargaining between buyers and sellers. This is considered by [Yavaş and Yang \(1995\)](#). In their model, a high list price will result in a higher sales price if an agreement is reached, but it will reduce the probability of sale. Among the empirical findings of this paper is that an overlisted house, one with a high list price relative to a hedonic estimate, will sell for more but take longer to do so.

The second impact of asking price is to encourage search. Yavaş and Yang did not consider this effect. [Horowitz \(1992\)](#) did, in a model where the list price is a ceiling and visits are encouraged by a low list price. The latter is not given microfoundation. His structural estimation of this model shows that reductions of list price are not likely to do much, which he interprets as explaining the empirical regularity that list price is infrequently adjusted. [Merlo et al. \(2013\)](#) made the point that this result is an inherent feature of a stationary search framework such as the one that Horowitz employs. As discussed below, Merlo et al. provided an explanation for the stickiness of list price in a nonstationary framework.

13.5.2 Asking price as a commitment

[Chen and Rosenthal \(1996a,b\)](#) presented models where asking price directs search by operating as a commitment mechanism. The simplest version of these models is as follows. A seller sets an asking price. Buyers choose to visit, and those that do learn their idiosyncratic match values. The seller also learns the match value (as in other games where the after-meeting price is determined by bargaining). The seller sets a take it or leave it price. This price is assumed to be constrained by the asking price.

The key result is that search can be encouraged by setting a low asking price. The result extends but in weaker form when the bargaining power is split between the buyer and seller. If the buyers have enough bargaining power, then a commitment through the asking price is not needed to encourage search.²¹

Instead of employing the standard Nash approach to bargaining, [Arnold \(1999\)](#) employed a [Rubinstein \(1982\)](#) bargaining game for the determination of prices. In this subgame, the asking price is, as in [Chen and Rosenthal](#), a ceiling on price. This generates a trade-off similar to [Chen and Rosenthal](#), with a lower asking price attracting additional visitors but resulting in a lower sales price as a negotiation outcome. The key difference in [Arnold](#) is that discount rates and outside options impact the transaction. In [Arnold](#), as in [Chen and Rosenthal \(1996a,b\)](#) and [Green and Vandell \(1998\)](#), asking price is modeled as a ceiling. But the asking price is not really a strict commitment. How does it matter in this case?

13.5.3 Asking price as a partial commitment

Although asking price is not a binding ceiling, it does seem to matter. One way to see this is to note that although houses sell for both less and more than their asking prices, a notable share of housing transactions ends with the price exactly equal to the asking price ([Han and Strange, 2014](#)). In other words, asking price is accepted at least some of the time. This is documented although not commented on in several waves of surveys by [Case and Shiller \(1988, 2003\)](#). [Han and Strange \(2014\)](#) provide further documentation using more than two decades of surveys by the NAR and more recent surveys in one market. So although it would be incorrect to model a home seller's asking price as a simple posted price (as with a good), it would also be incorrect to see the asking price as being meaningless.

This leads to two questions, one theoretical and one empirical. The theoretical question is: how can one rationalize in a fully specified equilibrium model how asking price can direct search even though it is neither a posted price nor a ceiling? The empirical question follows naturally: how does this kind of partial commitment impact search and housing market outcomes in various stages of the real estate cycle?

[Han and Strange \(2014\)](#) present a model showing how asking price can impact housing transactions even though one observes sales prices above, below, and equal to asking price. The heart of the model is as follows. Suppose that buyers have a two-point distribution on the match utility from a particular house. Suppose the asking price is a partial commitment in the following sense. When all buyers have match utility less than asking price, then sales price is determined in a Nash bargain. When two or more buyers have match utility greater than asking price, then sales price is determined as in an auction (we will discuss auctions further below). When exactly one buyer prefers the asking price to

²¹ See [Lester et al. \(2013\)](#) for a demonstration that a ceiling asking price is an efficient mechanism.

the price that would come from a Nash bargain, then the sales price equals the asking price. In this case, the seller will use the asking price to encourage visits. The seller's trade-off is between the surplus that is sacrificed in the acceptance case and the greater likelihood of the profitable bidding war case.

Han and Strange (2014) provide empirical evidence consistent with this role. The analysis is unusual in making use of data on buyer search behavior and not just on aggregate outcomes such as price and time-on-market. Several results emerge. First, a lower asking price increases the number of bidders on a house (a subset of the number of visitors). Second, asking price has a stronger negative relationship with search activity in a bust than in a boom. Third, the asking price performs this directing role even though housing is clearly not a posted-price market, with houses sometimes sold below list, sometimes above list, and sometimes at list.

13.5.4 Asking price as a signal

Albrecht et al. (2012) offered a directed search model that also has the realistic feature that sales price can be below or above asking price. They begin with a one-period model where all sellers are identical. Buyers choose which houses to visit, with the visit revealing to the buyer the idiosyncratic match value of the house. Buyers then bid, with this initial bid between the reservation price of the seller (assumed to be known to buyers) and the asking price. If no buyer bids at or above asking price, then the house is sold to the highest bidder at the bid price. If exactly one buyer bids at the asking price (which means that all others bid less by construction), then the house is sold at the asking price. If more than one bidder bids the asking price, then there is a second round that proceeds as an English auction. It is assumed that buyers make visit decisions without knowing how many other buyers might have chosen to visit a given house. In this homogeneous seller setting, it is possible for sales price to be below, at, or above asking price in equilibrium. A seller's revenues are independent of which of these marketing strategies is followed.²²

Now suppose sellers differ in motivation (reservation price). In this case, asking price can signal motivation. With two types of seller, we will have two asking prices, one higher for the less motivated sellers and one lower for the more motivated sellers. A more motivated seller will have more visits, giving a higher probability of sale (capturing liquidity in this model). The price conditional on sale will be lower, however. In both cases, the equilibrium is efficient, as in labor models.

13.5.5 Directed search and bargaining

Merlo and Ortalo-Magné (2004) documented the housing transaction using an unusual data set in a way that sheds considerable light on directed search. Their data cover fewer

²² See also Wang (2011) for another model of how list price can signal quality.

than 1000 transactions from three London area real estate boards during a time of market stability. However, the data include information on a number of features of the housing transaction that are usually not documented. The data record all reductions in asking price and the time of reduction. They also record all offers that are made. In the English institutional setting, this amounts to recording the entire history of negotiations, since sellers in England do not typically make counteroffers as they would in North America. So in addition to working with sales price and time-on-market, as is typical, Merlo and Ortalo-Magné were able to characterize the microstructure of the transactions in their sample.

Using these data, Merlo and Ortalo-Magné showed that a high list price is associated with a high sales price but a slower sale. List reductions are observed most frequently for houses that have not previously attracted much interest as measured by offers. Houses frequently sell to the first buyer who makes an offer, but one-third of matches that have led to offers fail to lead to a transaction.

This paper is an excellent example of the sort of impactful contribution that can be made using new and idiosyncratic data sources. Prior work typically used data that reported asking and sales price and also time-on-market for sellers. Such traditional data reported almost nothing about either search activity directly (i.e., numbers of visits) or the bargaining process (i.e., offers and counteroffers), especially in cases where a meeting did not result in a sale. And traditional data are almost completely silent on buyer search activity. There is a strong case for the creation of new data such as those used by Merlo and Ortalo-Magné or making use of new Internet-based data such as in [Piazzesi et al. \(2013\)](#).

13.5.6 Structural approaches to directed search

[Carrillo \(2012\)](#) specified and estimated a stationary equilibrium search model of the housing market. His model incorporates two important features of the real estate market: heterogeneity in the buyer's and seller's motivation to trade and the directing role of the asking price. Note that in his model, houses never sell above their asking prices because by assumption, there is no *ex post* competition among buyers. The model is estimated using the MLS data for real estate transactions in Charlottesville City and Albemarle County, VA, in 2000–2002. He found that more than half of the buyer's home valuation is gathered at the time when a listing is viewed. In addition, the model is used in two applications. First, it is used to evaluate the information content of visits and to determine the impact of previsit information, issues that are obviously central to search theory. Second, he also considered the impact of the structure of agent compensation. Counterfactual experiments show that the asking price and sales price are both reduced when there is additional online information about listings or when the commission rates are reduced.

Merlo et al. (2013) took a structural approach to solve a series of decision problems that a home seller faces: which price to list the house at initially, how to revise this price over time, whether or not to accept the offers that are received, and whether to withdraw the house if offers are not sufficiently attractive. They formulated these decisions in a finite-horizon, dynamic programming model, taking as given the seller's expectation about the probability a potential buyer arrives and makes an initial offer, the probability she will make additional offers if being rejected, and the level of the offer prices. Since they calibrated the arrival and negotiation process using UK data (as in Merlo and Ortalo-Magné, 2004), they adopted the UK institutions. An important feature is that there are no seller counteroffers, meaning that negotiation involves some number of buyer offers, each of which the seller must choose to take or to leave. These empirically fitted automata are then employed to estimate a seller's optimal dynamic sales strategy.

A number of interesting results emerge. First, small menu costs result in rare adjustments of asking price (contrary to Salant's theory, where asking price is adjusted every period). Reservation price, however, does adjust over the sales period. The kind of structural approach taken in this paper seems like a very promising path toward a deeper understanding of the micro behavior of house buyers and sellers. Of course, this estimation requires data that are much richer than are usually available. It includes not just the usual outcome variables (asking price, sales price, and time-on-market) but a number of process variables (individual offers made by buyers and asking price revisions).

While both Carrillo (2012) and Merlo et al. (2013) explicitly allowed buyers to direct their response to sellers' asking price, their approach on modeling the bargaining process is highly stylized. In the former case, trade occurs either at the asking price or at the seller's reservation value; in the latter case, buyers are treated as bidding automata and the offer process is one-sided. In both cases, sellers interact with only one buyer at a time. This prevents *ex post* competition among buyers for a seller's house—a point that we will come back to in Section 13.6 where we discuss auctions.

13.5.7 Directed search and booms and busts

In Section 13.3, we discussed some quantitative models of random matching. Díaz and Jerez (2013) computed a quantitative model of directed search. It allows the determination of the magnitude of the effect of housing search frictions on the housing cycle. This calibration exercise includes a wide range of general equilibrium effects. The results are consistent with housing frictions amplifying shocks and producing greater volatility.

In a similar vein, Caplin and Leahy (2011) considered how a directed search model can generate the key qualitative characteristics of housing market dynamics. They noted that a basic search model can generate volatility, a price–volume correlation, and a negative correlation of inventory and price. A basic search model fails, however, to generate the positive autocorrelation in price found by Case and Shiller (1989). This requires a

modification, such as information frictions where agents do not know the state of the market. In Caplin and Leahy's analysis, bargaining power changes endogenously over cycle (unlike other models, such as [Wheaton, 1990](#)).

13.5.8 Housing search and labor markets

We have thus far focused on housing frictions. Housing markets are, of course, related to labor markets, and labor markets have frictions too. This raises the possibility that the two sorts of frictions interact with each other. The idea that homeownership might be negatively related to labor market flexibility is often attributed to [Oswald \(1997\)](#). This empirical relationship can be given a search-theoretic foundation. Owners move less than renters, implying that changes in the location of a worker's ideal job may result less frequently in relocation for owners than for renters. In some situation, this may result in labor market mismatch.

Evidence on the Oswald Hypothesis has been mixed. [Coulson and Fisher \(2002\)](#) showed that in cross section, owners do not have inferior labor market outcomes to renters in an OLS framework. In [Coulson and Fisher \(2009\)](#), the result is shown to extend for the most part to instrumental variation. They attributed the absence of an Oswald effect to the adjustments of firms. On the other hand, [Munch et al. \(2008\)](#) established a negative relationship between homeownership and job mobility in Denmark. Likewise, [Battu et al. \(2008\)](#) showed that both ownership and residence in public rental housing can discourage mobility. The latter confers a kind of tenure whose security and duration approach that of ownership.

[Head and Lloyd-Ellis \(2012\)](#) offered a resolution. They presented a directed search model of the interaction between labor and directed housing search. The illiquidity of housing produces frictions in the labor market as well, with the rate of accepting out-of-town job offers being lower than the rate for job changes that would not require a home sale. A calibration shows the effect of homeownership to be small with a relatively small rate of unemployment. However, when unemployment is large, housing illiquidity has an economically significant effect.

13.6. AUCTIONS

We have thus far dealt with sequential search where buyers and sellers are matched one to one. In auctions, the matching is multiple. In this case, the bargaining that takes place is no longer a one-to-one process of negotiation. An auction may sometimes involve many buyers bidding for a single house. It may instead involve many buyers bidding for many housing units (typically condominiums). This section will deal with these auctions.

Our first point on this topic is that auctions happen. There are English auctions for houses in Australia ([Lusht, 1996](#)). There are auctions with sealed bid characteristics in Scotland ([Pryce, 2011](#)). And there are multiple object auctions as well ([Ashenfelter](#)

and Genesove, 1992). Auctions take place in both the single-family residential and the multifamily markets. They also take place in nonresidential markets and also in the land market (Quan, 1994, discussed many areas where auctions have been used).

In fact, auction-like transactions have become considerably more common during the great boom of the 2000s, and they have retained much popularity during the bust. We are referring here to the so-called bidding wars, where multiple buyers compete against each other for a property, typically very soon after listing.²³ While standard data sources almost never document the bidding for a house, one can infer the presence of competing bids when a house sells for more than its asking price. Han and Strange (2013) showed that the fraction of above-list prices was a fairly constant 3–5% prior to 2000. It grew to a national average of roughly 15% in 2005 before by about half in 2008. The low figure is consistent with Merlo and Ortalo-Magné (2004), whose unusual data do report bids, while the high figure is consistent more recent data reported in Han and Strange (2013). In some markets, bidding wars became much more common (roughly one-third of sales). These transactions involve bidding in an informal auction. An auction is not guaranteed, however, and many house sellers who hope for auctions must ultimately sell through the traditional sequential mechanism. This has led to the phenomenon of “backup offers” (discussed in Ashenfelter and Genesove, 1992; Quan, 2002; Ooi et al., 2006). In this situation, the distinction between auction and sequential sale is at least somewhat fuzzy.

The existence of auctions leads naturally to two important and related questions: How do auctions perform relative to more traditional sequential sales mechanisms? What determines when an auction is chosen or whether one arises naturally?

13.6.1 Theoretical issues

There is a vast literature on auctions as a general phenomenon. Krishna (2009) is a comprehensive reference. A buyer’s participation in an auction is central to the outcome it delivers. Such participation is costly. McAfee (1993) and Peters and Severinov (1997) are classic references on competing auctions where the competition takes place through posted reserve prices. Albrecht et al. (2012) considered efficiency in this setting. Wang (1993) examined the related issue, the relative performance of auctions versus posted-price sale in a private value setting. Buyers arrive randomly with random match values. If the seller chooses to post a price, then a transaction will occur when a buyer arrives with a sufficiently large valuation. An auction, in contrast, involves choosing a critical time, with all buyers arriving before that time allowed to participate in the auction. The seller incurs costs under both mechanisms while waiting for a sale. The seller also incurs a fixed cost of holding an auction. Wang showed that an auction generates greater revenue when this auction cost is zero. Even with a positive auction cost, an

²³ Pryce and Gibb (2006) showed that booms are associated with an increase in the number of buyers competing for a house in Scotland, where the system more closely resembles an actual auction.

auction generates greater revenue when the marginal revenue curves defined by buyer valuations are sufficiently dispersed.

Bulow and Klemperer (2009) considered why sellers might prefer auctions. They noted that the simple answer that auctions deliver competition and thus high prices is not entirely satisfactory, since a sequential sales mechanism also confronts a buyer with competition from future rival buyers who may arise. In Bulow and Klemperer's model, sequential search is efficient in the sense that it encourages buyer visits (participation) precisely when such visits are most valuable: when prior buyers had low valuations. Auctions, in contrast, are inefficient, with many buyers making participation decisions without information regarding other buyers' match values. Because of this, auctions give more dispersion in outcomes. This dispersion encourages buyers to participate, which in turn is why in most situations, sellers prefer auctions. The key exception noted by Bulow and Klemperer is when buyers cannot issue credible "jump bids" that preempt further search activity. In this case, it is possible that a sequential process would give a higher price.

The literature on real estate auctions is small, especially on the theoretical side. See Quan (1994) for a discussion of the implications of general research on auctions for real estate. Adams et al. (1992) modeled real estate transactions as a "slow Dutch auction." In this setting, when a house seller faces a stationary environment, it is optimal for the seller to maintain a fixed asking price, and it is never optimal for the seller to terminate the process by calling an auction. Mayer (1995) pointed out that this result requires stationarity. In his model, auctions deliver rapid transactions but at a discount. In a down market, the auction discount will be steeper. With seasonality (as in Salant, 1991), the optimal price may fall over the marketing period and an auction may deliver a higher price than a sequential mechanism. Quan (2002) presented a model with an additional element. Buyers, like sellers, can differ in their taste for a rapid transaction. Buyers are willing to pay more to avoid a lengthy period of search. Sellers are willing to accept less.

13.6.2 Empirics

13.6.2.1 Auctions and outcomes

A number of papers have considered the empirical relationship between the use of an auction instead of a sequential sale mechanism and the price of a house. Lusht (1996) considered this issue in the Australian setting where auctions are common. Controlling for a house's characteristics, he found that houses sell for more at auctions. Of course, the choice to auction is endogenous, and this could produce biased estimates of the auction coefficient if unobserved house characteristics that buyers value are correlated with the decision to auction. Quan (2002) found a premium for auctioned properties even using a method-of-moments approach to control for unobserved heterogeneity. Mayer (1998), in contrast, found a discount for auctions using a repeat sales estimator to control for endogenous selection into auctions. This is explained as reflecting the seller trade-off

between price and liquidity. [Chow et al. \(2014\)](#) help to explain these disparate results by showing that the difference between price under auction and that under negotiated sales depends on property type and market condition. They show that the auction premium is larger for more homogeneous properties and in booms.

All of these papers consider the impact of auctions within small data sets. [Campbell et al. \(2011\)](#) considered the effect of a “forced sale” on price in a very large data set. Their goal in this regard is to use quasi-experimental variation in the type of sale—forced or not—in order to assess the liquidity discount.²⁴ They did this by obtaining data on deaths and delinquencies, both circumstances that transmit the property to another owner who is likely to be mismatched with the housing unit and therefore motivated to sell. They found a forced sale discount of 3–7% across all forcing events in OLS models. However, there is reason to be concerned that the characteristics of the housing unit are related to the circumstance forcing the sale. An older household may not maintain a house in the same way that a younger one would, for instance. They addressed this concern by estimating models for different forcing circumstances, for different ages of owner, for different types of property, and separately for the structure and land components of property value. They found evidence consistent with the discount associated with death being “primarily” associated with unobserved housing characteristics. They did not find this for mortgage foreclosures or bankruptcies. In these cases, they concluded that prices of forced sales are indeed lower.

13.6.2.2 *Bidding wars*

As noted above, formal auctions remain rare for nondistressed single-family houses in North America. They are also uncommon in the nondistressed multifamily residential market. However, informal auctions do take place. With “bidding wars,” a house (typically newly listed) is sold in a process of competitive bids made by rival buyers. The exact institutions are variable. In Toronto, written bids are presented at a particular time, typically accompanied by a presentation by the buyer’s agent. After all bids, the seller typically involves some or all of the initial bidders to bid again. This can be repeated. Buyers usually are told the number of other bidders before they submit the bids. While a bidding war clearly is a sort of auction, it is not one of the standard forms. In Washington, DC, the process is more formal. Bids are submitted with escalators specifying how high the bidder will go. Buyers usually do not learn of the number of other bidders or the bids. These bidding wars are much more common than they used to be. As noted above, the share has more than doubled between the late 1990s and the early stages of the housing recovery. The fraction of sales taking this form varies significantly across space.

²⁴ An equally important goal is to understand the effects of foreclosures, a particular type of forced sale, on housing markets. This is not our central focus in this chapter.

Han and Strange (2013) considered the determinants of bidding wars. The analysis shows that the share of bidding wars is procyclical. There are also more bidding wars with greater Internet adoption. Sales with price greater than list are shown to occur early in a house's sales period. This does not, of course, mean that sellers can use informal auctions like this to achieve both high prices and low seller time-on-market.

All of this makes clear that a seller does not typically decide between auction and sequential search. The decision is made by the market. The seller may influence this by strategically marketing the house. A credible commitment to a low list price may perform this role. Some houses may attract enough interest to sell at a high price and soon. Others may not. It is likely that at least some of the premium that is associated with auction and auction-like transaction mechanisms reflects this sort of unobserved heterogeneity.

Or it may reflect something quite different. Ashenfelter and Genesove (1992) considered the possibility of behavioral anomalies influencing real estate auctions. This possibility is clearly present in popular discussions of bidding wars, with folklore suggesting that a low listing price to bring people to the table can result in a high sales price as bidders throw caution to the wind in the ensuing auction.

To consider this possibility, Ashenfelter and Genesove (1992) worked with data from a small multiobject "pooled" auction. In such an auction, the highest bidder chooses a property first. In such a model, price will be lower for later units since they will be inferior in a revealed preference sense. Since units are resold after bargaining, one can address the omission of characteristics by looking at how private sale prices relate to the time of sale within an auction. Their results are surprising. They found prices of condominiums sold at auction were 13% higher than prices for identical units sold through postauction bargaining. These units were actually the same properties, rather than being identical in the hedonic sense of having identical observed attributes. This is interpreted as a possible instance of winner's curse. If buyers are behaving irrationally in this way, then sellers could profit from choosing auctions. Of course, it is unclear how such a strategy would hold up when enough sellers used it and buyers became aware of it.

The increasing popularity of bidding wars in recent years also permits an opportunity for researchers to empirically measure the thinness of the housing market. Even a casual acquaintance with the market, such as most people's experience of buying or selling a home, would attest to the essential thinness of the housing market—the variability in the match quality between buyers and sellers and the inability of buyers to find suitable sellers and vice versa without making costly efforts, which underlie all search and matching models. However, the thinness itself, despite its intuitive appeal and fundamental importance, is neither easily defined nor readily measured.

Employing a recent survey that collects information about home search, bargaining, and bidding process in a large North American metropolitan area in 2005–2009, Genesove and Han (2012b) estimated market thinness in two ways. First, controlling

for observed and unobserved heterogeneity among housing units, doubling the number of bidders increases the sales price by 2.4%. This effect is statistically significant. This is contrary to the hypothesis of a thick market in which buyers' valuations are homogenous and so the sale price is invariant to the number of bidders. Second, the spread of bidders' valuations for a given house, measured by the standard deviation of the underlying distribution, falls within the range of 4–5% of home value. Intuitively, if homes are not very different one from another, one buyer will evaluate a given house pretty much the same as another. The enormous estimated buyer valuation variance for a given home provides a clear demonstration of how thin real estate markets are.

13.7. REAL ESTATE BROKERS: FUNDAMENTALS

A large fraction of search, matching, and bargaining in the housing market is facilitated by real estate brokers and agents.²⁵ This section begins to consider the role of intermediation in real estate transactions. It describes the activities of real estate brokers, and it discusses why they play the role that they do. Later sections consider the nature of competition and entry in the brokerage industry, the incentive issues that impact broker performance, and the efficiency of resource allocation in this sector. In other words, the chapter now moves on to consider the intermediation factors discussed in [Figures 13.1 and 13.3](#).

13.7.1 The activities of real estate brokers

One key difference between intermediaries in the financial and commodity markets and intermediaries in the real estate markets is that the former are dominated by market makers, while the latter are dominated by matchmakers. According to [Yavaş \(1994\)](#), a market maker sets an asking price and a bid price at which she buys and sells for her own account. Examples include specialists in the stock market and dealers in the used-car market. In contrast, match makers create meetings between the buyers and the sellers rather than participating in the trade themselves. One interesting phenomenon is that real estate markets typically only have the second type of intermediary, but not the first.²⁶ [Anglin and Arnott \(1991\)](#) attributed the absence of dealers in the real estate market to high inventory cost and the risk associated with houses. Using a search-theoretic model, [Yavaş \(1992\)](#) showed that the search costs associated with

²⁵ As noted in [Section 13.1](#), we use the terms “broker” and “agent” interchangeably.

²⁶ There are, however, exceptions. For instance, there are “flippers,” who hold housing for speculative reasons, as documented by [Bayer et al. \(2011\)](#). As of now, this class of intermediaries is not commonly involved in housing transactions, possibly because of transfer taxes, holding costs, and other related barriers to this form of intermediation. Furthermore, there seem to be cases when a broker will buy a house and then later return the house to the market.

finding trading partners help to explain the dominance of brokers over dealers in the housing market.

More specifically, real estate agents and brokers are licensed professionals whose main job is to match a home seller with a homebuyer. Together, they provide a bundle of services to buyers and sellers. An agent working with buyers is often referred to as the “cooperating agent” or “selling agent.” Cooperating agents typically attempt to find houses that match buyers’ tastes, show buyers prospective homes, advise them in making offers, and provide assistance in the negotiation process. An agent working with sellers is often referred to as the “listing agent.” The listing agent helps sellers list the house on the MLS, assists sellers in staging and marketing the house, advises sellers on the listing price, helps sellers evaluate offers and formulate counteroffers, helps negotiate directly with the buyer or the buyer’s agent, and provides assistance in closing a transaction. In North America, the listing agent is commonly contractually granted the exclusive right to sell. Typically, the cooperating agent is legally considered to be a subagent of the listing agent.²⁷

It is common for agents to receive their compensation as a fixed percentage of the sales price of a home. In North America, it is typical for the commission to be 5–6% of the sale price. The listing agent shares equally with the cooperating agent who brings the buyer to close the deal. Both agents further split commission fees with their affiliated brokerage offices, in return for the brand value and for the supporting services that brokerage offices provide.

It is important to recognize that the North American situation is not universal. In England, a commission of 1–2% is more common, for instance, and there is only one-sided seller representation. Furthermore, while contracts in North America typically specify exclusive representation by a single broker, sellers in England can choose to pay a higher commission, perhaps 2.5–3%, and be represented by multiple agents. In Vancouver, the contract is not even linear. The typical real estate commission is 7% of the first \$100,000 and a lower percentage (usually 2–3.5%) of the rest of the sales price.²⁸

The compensation structure and the nature of competition in the real estate broker industry have been the subject of extensive prior research, and excellent surveys already exist for the interested reader (see, e.g., recent surveys by [Yavaş, 1994](#); [Benjamin et al., 2000](#); [Miceli et al., 2007](#); [Zietz and Sirmans, 2011](#)). For our purposes here, we will not reproduce this discussion but rather focus on the more narrow, but crucial, topics of search intermediation, competition and efficiency, and information economics.

²⁷ In commercial real estate, there is typically only one agent.

²⁸ A lower commission for the marginal dollar of house value makes the contract even more low-powered than with linearity. This is unlikely to improve agent incentives.

13.7.2 Why do real estate brokers exist?

According to a survey conducted by the NAR in 2005, 84% of single-family house sales in North America are facilitated by real estate brokerage. However, as illustrated in Sections 13.2–13.6, neither buyers nor sellers must use real estate agents; counterparties can approach each other directly in housing markets. Thus, the most basic question regarding real estate brokers is: why do they exist?

The consensus of the literature is that real estate brokerage emerges mainly due to imperfect information. For example, [Yinger \(1981\)](#) stressed the importance of information in a search-theoretic framework. He discussed three types of uncertainty. There is uncertainty about the number of buyers, the number of listings, and the number of matches. In this setting, Yinger showed that the existence of MLS improves outcomes and thus confers a competitive advantage on member brokers. [Wu and Colwell \(1986\)](#) extended Yinger's model to incorporate the behavior of buyers and sellers. [Yavaş \(1992\)](#) further extended this line of literature by incorporating the bargaining stage into the model. In an elegant search-theoretic model, Yavaş showed that a search economy without brokers involves two sources of inefficiency. First, as with Yinger, brokers are better informed. Second, as with many search models, there are positive externalities. An increase in the search efforts of either a buyer or a seller increases the probability of a match, hence the payoff for the counterparty. Because neither takes into consideration the effects of own search on the additional gains to the counterparty, this positive externality results in less search activity than would be in the joint interest of the two agents. These inefficiencies create a situation where real estate brokers can potentially reduce the uncertainty and internalize some of the externalities in return for some profits. Thus, one reason for real estate brokers to exist is that they create value by resolving information problems.²⁹

The informational need for an intermediary is further exacerbated by a set of intrinsic characteristics of housing: necessity, heterogeneity, indivisibility, complexity, high stakes, and transaction costs ([Arnott, 1987](#)). These characteristics interact to make it costly for buyers and sellers to find the right trading partners and to reach agreements on house prices. In particular, the multidimensional heterogeneity of a house, combined with the multidimensional heterogeneity of buyer tastes, renders the housing market thin. In a thin housing market, even with many house units available for sale, it can be difficult for a buyer to find a house that matches his preference. Moreover, unlike participants in financial securities markets, most homebuyers and sellers have very limited experience in trading a home. Households typically buy and sell a small number of houses over a lifetime because of the high transaction cost associated with buying and selling a house

²⁹ See [Kurlat and Stroebe \(2014\)](#) for strong evidence that there exist information asymmetries in real estate. While their analysis does not focus primarily on the role of real estate agents, some of their analysis is based on the idea that brokers operate with less uncertainty.

(Haurin and Gill, 2002). In addition, given that house is the largest financial asset in a typical household's portfolios (Tracy and Schneider, 2001) and given that house price risk is not readily diversifiable (Caplin et al., 1997), the cost of a mismatch in the housing market tends to be particularly large.

Together, these features imply that search frictions are substantial. With their information advantage relative to homebuyers and sellers, real estate brokers can potentially improve the function of the market by dispersing information and creating a more coordinated matching scheme. Thus, a natural question to ask is what gives real estate brokers better information? In what follows, we discuss two potential sources.

The first is their superior access to MLS, which gives brokers better information. As Stigler (1961) had shown, the acquisition of information about the state of the market can absorb significant resources. This is particularly so for housing markets given the frictions described above. For any given housing transaction, there are advantages to having a centralized depository of information that summarizes offers to sell (listings).³⁰ Such information helps buyers to expedite the process of search and helps sellers to determine an initial asking price. By actively collecting and dispersing such information through MLS arrangement, real estate brokers can potentially reduce the frictions in the housing market and ameliorate the problem of imperfect information, permitting a more efficient search and bargaining process. As noted above, the impact of MLSs on the search strategy of buyers and sellers and brokers has been analyzed in detail by Yinger (1981) and Wu and Colwell (1986).

Of course, this situation has changed significantly with the spread of the Internet. This allows buyers to see active listings on the MLS. Furthermore, there are online resources such as Trulia and Zillow that can also help participants better understand the state of the market. However, it should be noted that brokers continue to have exclusive access to data on historical completed transactions. These data include information on sales price and time-on-market of previously completed transactions, rather than just the list price that is reported on an active listing. Thus, even as of this writing, the MLS confers a significant information advantage to brokers.

The second source of the real estate brokers' information advantage comes from the experience gained through their professional activities. Unlike a typical homebuyer or seller who is involved in a house transaction only a few times, an established real estate agent has been involved in many real estate transactions. This experience helps brokers build expertise in understanding market conditions and handling the financial and legal complexities involved in completing a real estate transaction. This expertise is valuable for homebuyers and sellers as the high stakes associated with risky housing transactions mean that mismatch can potentially be very costly. Quite naturally, the value of real estate

³⁰ An interesting question is why there is not a centralized depository of information that summarizes a list of potential buyers.

brokers should be particularly large for unsophisticated buyers and sellers with limited knowledge and experience of housing markets. Consistent with this, Benjamin et al. (2007) found that brokers are more likely to be employed by those with less knowledge about housing transactions.

The information benefits associated with real estate brokerage do not, however, come without cost. First, the commission fees that brokers charge create a spread between buying and selling prices. This could potentially prevent mutually beneficial transactions from taking place.³¹ It is parallel to what has been called excessive intermediation in other financial markets. In a market without agents, a seller and a buyer would trade with each other only if the latter's valuation of the house exceeds the former's. With agents, they will trade only if the difference between their valuations exceeds the commission fees that they have to pay the agent, which creates an additional friction. Thus, whether brokers can enhance welfare depends on whether the transaction cost they impose is lower than the benefit they bring by economizing on search costs. Jud and Frew (1986) and Yavaş (1992) showed that the seller receives a higher price when employing a broker, but the increase in price is less than the commission. This seems to suggest that the spread, the brokerage commission, might be large enough to be a significant barrier to transactions. However, brokerage may also improve liquidity and match quality while allowing buyers and sellers to choose lower search intensities and thus incur lower search cost.³² The welfare economics of the commission spread are, thus, quite complicated, depending on commission fees, search costs, and matching technology.

Second, with a market structure characterized by weak price competition and free entry, there will be a tendency for excessive entry. The ensuing inefficiency can take the form of inefficient scale in that there are too many brokers. It could also take the form of a broker spending too much search effort to compete with other brokers for new listings. Such effort might have private value to a broker but no value to clients and other brokers and therefore can be considered to be socially wasteful. This point will be further illustrated in Section 13.8.

Third, the information advantage that real estate brokers possess relative to buyers and sellers also raises the possibility that these agents may behave strategically. A broker's interest in maximizing commission revenues could cause a series of incentive alignment problems, which could introduce noise into the home search process and lead to a substantial welfare loss for the clients they work for. We will elaborate on this point in Section 13.9.

³¹ It is typical for jurisdictions to impose taxes on transactions, which further increases the spread between the price a buyer pays and the price a seller receives. Dachis et al. (2011) showed that this discourages transactions. In a similar vein, weaker property rights in a developing economy would also result in a lower expected payoff from search and in less search and fewer transactions.

³² See Yavaş (1994) for a theoretical demonstration that equilibrium search intensities fall when brokers are employed.

13.8. COMPETITION IN THE RESIDENTIAL REAL ESTATE BROKERAGE INDUSTRY

The brokerage industry is the subject of vigorous policy debate. A key question in this debate is whether this market is truly a competitive one. This section considers competition in brokerage from both positive and normative perspectives.

13.8.1 Nature of competition

A report by the [National Association of Realtors \(2005\)](#), NAR, provides an economic analysis of the structure, conduct, and performance of the real estate brokerage industry. The report argues that the industry is fiercely competitive, evidenced by the large number of brokerage firms and agents, the low concentration ratio in the brokerage market, and the ease of entry into and exit from the profession. For example, according to [FTC \(2007\)](#), there are approximately 98,000 brokerage firms operating over 200,000 local offices in the United States. These offices provide potential employment for approximately 2.5 million real estate licensees. The NAR further reports that in 2004, the top 10 brokerage firms have a combined market share of 9.1%, the top 20 firms have a share of 10.9%, the top 100 firms have a share of 17%, the top 500 firms have a share of 26.6%, and 96% of brokerage offices employ 10 or fewer real estate agents.

A related strand of the literature argues that the competitive nature of the brokerage industry is also evidenced by the absence of cost inefficiencies. For example, papers by [Anderson et al. \(2000a,b,c\)](#) look at X-efficiencies and conclude that the performance of real estate brokerage is economically efficient. Specifically, firms are more likely to fail to maximize profits than to fail to minimize costs.

Not all researchers are convinced about the competitive nature of real estate brokerage. Most of the disputes related to this issue have centered around three structural features of the industry: low entry barriers, limited product differentiation, and lack of price competition (see [Figure 13.4](#)). Below, we will discuss each of these features.

13.8.1.1 Entry

There are three relevant entry issues. The first is the entry of individual agents and brokers. The second is the entry of brokerage firms. The third is the entry of new business models for brokerage.

At the individual level, entry is relatively easy. The requirements for becoming a real estate agent seem to be minimal compared to other professions. A 1983 FTC Staff report on the real estate brokerage industry observed that “the nearly universal opinion is that there are no significant barriers to entry, if entry is construed as gaining a license in order to practice.” The ease of entry was further confirmed in a more recent DOJ and FTC Report in 2007. In particular, for an agent, the exam to obtain a license is the only barrier, and the requirements to pass the exam are limited. Moreover, at any point in time, there

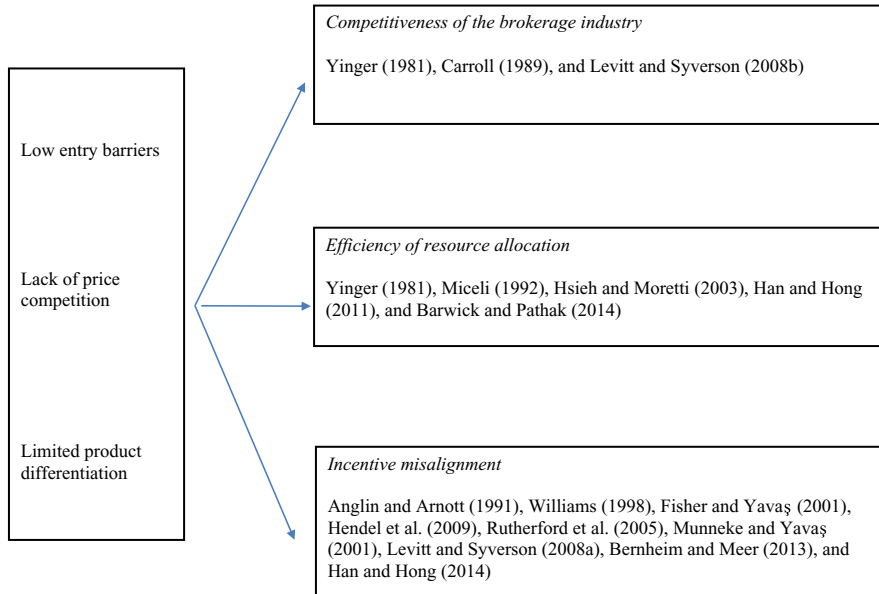


Figure 13.4 The Industrial Organization of the Real Estate Brokerage Industry. *Note: This figure lists three key characteristics of the brokerage industry and three strands of the literature that have emerged. The papers listed are examples only; see the text for a more complete review.*

are a large number of licensed but inactive agents who are presumably ready to become active when there is a profitable opportunity. According to [Hsieh and Moretti \(2003\)](#), about 20% of licensed agents are inactive. In most jurisdictions, broker entry requires experience as an agent and additional examinations.

Entry as a brokerage firm is more difficult. At a minimum, an entrant that wants to establish a brokerage firm must hire or become a licensed broker. Additionally, an entering brokerage firm may require an agent workforce, office space, office staff, and advertising. To establish a name recognition, a brokerage often has to pay a certain amount of franchising fees to be affiliated with a national franchise (e.g., RE/MAX) or to invest significantly to establish its own brand value.

While entry barriers at the agent and brokerage levels seem modest, entry barriers for new business models that challenge traditional full-service agency are considerably higher. This is because MLS membership is limited to licensed real estate brokers and agents that conduct business in particular manner, thereby limiting competition in price and in product variety. For example, the [FTC Report \(2007\)](#) discusses rules that discriminate against brokers who enter into exclusive-agency listing agreements and rules that discriminate against brokerage that takes place through virtual office websites (VOWs). The latter allow brokers to withhold their clients' listings from VOW brokers

by means of an “opt out” and limit competition from real estate brokers using innovative business models and the Internet to offer better service to their clients. These types of anticompetitive rules effectively erect entry barriers for new business models, thereby limiting price and service competition in this industry.

13.8.1.2 Product differentiation

Compared to other service industries, there is limited differentiation in the services that brokers provide. Traditional real estate brokers tend to provide a full package of services, including helping buyers perform MLS searches for homes, accompanying them on visits, helping sellers stage homes, making the house available for viewing, advertising, and setting the initial asking price. Despite the great potential for unbundling this full package of services, there is actually quite a lot of similarity in the scope of services delivered by traditional real estate agents. As discussed above, this is in part a consequence of market power associated with the MLS. In addition, some jurisdictions prevent licensed brokers from unbundling MLS listings from other services. For example, in some states, there are so-called minimum service requirements that mandate that a broker must perform a full package of services for a client. As of 2007, eight states in the United States had such law in practice (Bernheim and Meer, 2013).

There is contentious debate as to whether the services must be bundled. There is no technological reason for bundling, and standard economic arguments suggest that allowing choice from a menu of service options would have superior welfare properties. The feasibility of unbundling seems to have increased with the spread of the Internet. For instance, discount brokers will typically provide only one service, the listing of a seller's property on the MLS. They do not provide advertising or advice or any of the other elements of a full-service broker's package. The greater availability of information online presumably substitutes for these services. The rebuttal offered by brokers is that minimum service requirements ensure quality. In Section 13.9, we will review the recent literature that empirically examines the effects on selling price and time-on-market of discount brokerage and other business models.

Despite the limited degree of differentiation in service variety, real estate agents do tend to differentiate themselves by housing market segment. Almost all brokers specialize geographically. This is partly because housing markets are highly local, and as a result, the human capital that brokers accumulate over time, such as local experience, connections, and licensing requirements, is attached to the local market too. Beyond geographic specialization, some brokers further specialize by price segment, housing type, and the ethnic and social characteristics of clients. Others choose to specialize in listing or selling activities. For example, Turnbull and Dombrow (2007) showed that agents who specialize in listing properties obtain higher sales price for sellers they represent, while those who specialize in selling obtain lower prices for buyers they represent.

13.8.1.3 Price competition

An ongoing theme in research on the real estate brokerage industry is the extent to which pricing is competitive. Using the buyer-side commission rate, a number of papers show that commission rates are market-driven. For example, using 1983–1987 MLS data in Knoxville, Tennessee, [Goolsby and Childs \(1988\)](#) found that there is competition in the commission rate charged by real estate brokerage firms. In particular, firms are more willing to accept a lower commission rate when houses are newer or of higher value.³³ A more recent study by [Schnare and Kulick \(2009\)](#) uses MLS data for several metropolitan areas over the period 2000 to mid-2007. They found that buyer-side commission rates are strongly correlated with supply and demand side variables such as list price, number of transactions, number of agents, and limited-service listing agents. Their results suggest that commission rates are market-driven and competitive. To the extent that much of the negotiation of the commission rate occurs privately between home sellers and brokers and does not get reported on the local MLS, these findings may underestimate the true variations in the commission.

However, many observers believe that commission rate is still quite inflexible both across markets and over time. [Owen \(1977\)](#) provided evidence of this inflexibility. Later, using a nationally representative survey of home sellers and MLS data from Boston, Minneapolis, Los Angeles, and Seattle, the Federal Trade Commission's ([FTC, 1983](#)) report offers further evidence that commission rates were remarkably uniform in the late 1970s and early 1980s. Using the Consumer Expenditure Survey (CEX) from 1980 to 1998, [Hsieh and Moretti \(2003\)](#) showed that the commission rate continues to be fixed around 6% between 1980 and 1998. Using the statistics from the 2004 Real Trends Brokerage Performance Report, [Han and Hong \(2011\)](#) showed that the commission rate was around 5.1% in 2002–2003, with negligible changes in the commission rates across regions. More recently, using transaction data for the Boston area from 1998 to 2007, [Barwick and Pathak \(2014\)](#) found that there is little variation in commissions over time despite the increased penetration of the Internet and new technologies.

Despite the conflicting evidence about the degree of inflexibility in commission rates, it has been generally acknowledged that real estate brokerage industry has been quite successful in preserving its relatively rigid commission fees. This is particularly so when one compares the slow changes in commission rates for real estate brokers under Internet diffusion with rapid changes in commission fees in other agent-based computer service industries, such as travel agencies, stock brokers, and automobile insurance ([Levitt and Syverson, 2008b](#)).³⁴

³³ Other studies along this line include [Carney \(1982\)](#), [Sirmans and Turnbull \(1997\)](#), [Sirmans et al. \(1991\)](#), and [Miceli \(1992\)](#).

³⁴ Examples of papers dealing with uniform commission rates include [Owen \(1977\)](#), [Carney \(1982\)](#), [Crockett \(1982\)](#), [Wachter \(1987\)](#), [Goolsby and Childs \(1988\)](#), [Miceli \(1992\)](#), [Williams \(1998\)](#), and [Arnold \(1999\)](#). Different conclusions on the uniformity of commission rates are found in [Sirmans and Turnbull \(1997\)](#).

This pronounced and puzzling feature of brokerage commissions, combined with the ease of entry and lack of product differentiation, has motivated three lines of research, as depicted in [Figure 13.4](#). The research deals with the competitiveness of the real estate brokerage industry, the efficiency implications for resource allocation, and the possible misalignment between the goals of the brokers and those of their clients. The previously introduced [Figure 13.3](#) presents a summary of selected literature regarding each of these intermediation factors. We will discuss research on each of these factors in [Sections 13.8.2, 13.8.3, and 13.9](#), respectively.

13.8.2 The “commission puzzle” and the competitiveness of brokerage

The relatively uniform commission rate across markets and over time has led to a long-running debate about the competitiveness of the real estate brokerage market. Some view the fixed commission rate as being consistent with competitive pricing. Others view it as an indicator of price discrimination supported by tacit collusion among brokers. In this section, we will present the theoretical work on both views and then discuss the related empirical evidence.

13.8.2.1 Theoretical work

Several economists argue that the fixed commission rate might be consistent with competitive pricing. Suppose that the marginal cost of selling a higher-priced unit is greater. In this case, a fixed commission could be consistent with a competitive equilibrium in which higher-priced homes receive more service and hence pay more commission fees.

In particular, some studies have investigated the competitive nature of brokerage commissions with a focus on understanding the possible gains of having a percentage commissions relative to a uniform fee across clients. For example, [Schroeter \(1987\)](#) demonstrated in the context of a queuing model that fixed-percentage commissions might be consistent with competition in the real estate brokerage industry. The essential prediction of his model is that, holding other things constant, brokers serving sellers of more expensive houses should serve smaller number of clients than do brokers representing sellers of less expensive ones. As a result, more expensive houses remain on the market for a shorter time than less expensive houses. [Zorn and Larsen \(1986\)](#) and [Carroll \(1989\)](#) showed that fixed-percentage commissions could emerge in a principal–agent setting. In particular, if home sellers cannot monitor real estate brokers’ efforts on their behalf, fixed-percentage brokerage commissions can promote welfare by giving brokers an incentive to tailor their services to their clients’ demands. Hence, there is a competitive equilibrium in which clients who value brokerage services more highly offer to pay larger commissions and consequently receive more selling effort from the broker. [Knoll \(1988\)](#) further showed how differences in commissions might be related to the exclusivity of the broker’s contract.

Not all researchers are convinced that broker pricing is competitively determined. In fact, many economists have substantial reservations about the competitive nature of the brokerage market. Instead, they attribute the relative inflexibility of the commission rates to collusion. The idea goes back to at least [Miller and Shedd \(1979\)](#), whose analysis seeks to explain the fact that commission rates remained stable or even increased in the 1970s while underlying home price rose dramatically. This explanation is based on the assumption that collusive pricing would result in stable commission rates over long periods of time, and it suggests that the brokerage industry is exploiting its joint monopoly power with increasing vigor. In his classic study of the real estate brokerage industry, [Yinger \(1981\)](#) viewed the fixed-percentage commission scheme as a clear indicator of price discrimination that would justify government intervention in this market. He argued that “the service provided to a seller by a broker is to find a buyer and to finalize the sale. The cost of this service, is at best, only marginally related to the value of house involved, but higher-income households, who of course buy more expensive houses, are willing to pay more for this service.” [Yavaş \(2001\)](#) presented a model where the fixed costs of brokerage (e.g., license fees; fees for local, state, and national realtor associations; continuing education expenses; and some of the office expenses) make it impossible to have competitive commission rates as the equilibrium outcome. He saw collusion as another possible explanation for the uniformity of commissions.³⁵

One issue that has puzzled observers is how a collusive equilibrium, if it exists, could be maintained in the industry. Standard economic theory suggests that the difficulty of sustaining collusion grows with the number of players. In 2005, there are about two million active real estate agents and brokers, associated with 100,000 firms ([White, 2006](#)). The large number of players, combined with the ease of entry into and exit from this profession and the low concentration ratio, makes the collusion particularly difficult to implement ([NAR Report, 2005](#)).

Researchers in this literature, however, have identified at least three institutional features that have the *potential* to facilitate collusion. First, the MLS provides the identification of price-cutting activity and discourages price-cutting through the incentives it offers ([Crockett, 1982](#)). Since the access to MLS is limited to licensed real estate brokers and agents who conduct business in a particular manner, it could provide a way for the local real estate board to enforce the collusion among the members, thereby limiting price competition. Moreover, under the MLS system, when an agent submits a listing to the service, the commission available for a cooperating broker is announced. A broker that solicits listings by offering lower commissions may find himself disadvantaged since agents from other firms may prefer to bring their buyer clients to visit listings

³⁵ [Anglin and Arnott \(1999\)](#) presented a formal general equilibrium of how the commission rate is determined. Their model reflects the externalities in the market and compares the equilibrium commission rate to the socially optimal rate.

with standard commission rates. Thus, those who have hypothesized that pricing in the real estate brokerage industry is collusive have suggested that the MLS in each market eliminates the threat of price-cutting by giving information about commissions to all brokers in the market, thus making it possible for brokers to punish price-cutters (FTC, 1983; Wachter, 1987). White (2006) concluded that “the ability of the collective members of a MLS to exclude rivals—especially if those rivals are ‘mavericks’ who are price-cutters, with respect to commissions—can be a powerful way of enforcing a high-fee structure and thus of maintaining the collective exercise of market power.”

Second, in North American residential markets, a complete real estate transaction requires coordination from both buyer-side agents and seller-side agents. Thus, it is possible for agents working on the traditional commission fee system to discipline the price-cutting agent by refusing to bring their buyers to see the latter’s property (Bartlett, 1981; Crockett, 1982). This mechanism could be quite effective in enforcing the collusion, even in the absence of the MLS (White, 2006). In particular, listing agents need cooperating agents to deliver buyers. Not only can traditional commission agents collude against price-cutters by steering their buyers away from such listings, but also they can use the same punishment on other traditional commission agents who cooperate with price-cutters. The common practice of giving one broker an exclusive right to sell could further create anticompetition (Braswell and Poe, 1992).

Furthermore, some policies that local real estate boards and NAR adopted could become barriers to competition. For example, in areas where VOW policies are in place, agents could withhold a large fraction of the houses listed for sale on the MLS from flat-fee agents’ VOWs. In addition, 10 states have bans on agents offering rebates to their clients (Han and Hong, 2011). These bans explicitly prevent real estate brokers from price-cutting and hence further facilitate collusion among real estate brokers.

13.8.2.2 Empirical work

Despite the extensive discussion about possible collusion among real estate brokers, there is only limited empirical evidence on this topic. The lack of the empirical work is probably due, in large part, to the following challenges. First, much of the commission data come from the MLS database, which only reports the buyer-side commissions but not the seller-side commissions. To the extent that much of the negotiation of the commissions occurs privately between home sellers and brokers, there is no direct evidence on the true variation of the commission rates. Second, discount brokerage business models typically do not report transactions through the MLS. Thus, MLS data alone are not sufficient for researchers to test whether any deviation from the standard business manner has been punished in a way that is consistent with collusion. Third, testing for competitive pricing would require measures of the broker’s capacity and skills, the sellers’ motivation, and the thinness of each segment of the housing market. Failure to adequately control for these factors would affect the interpretation of the

empirical findings. Finally, although a lot has been said about the possibility of collusion, no formal model of collusion in brokerage has been worked out yet. As [Carroll \(1989\)](#) pointed out, “in such a model, the uniform commission rate would emerge as the optimal outcome in a cooperative game. The model would explain how commission rate depends on market forces (such as home selling prices, brokers’ costs, and clients’ incomes) and institutional factors (such as informational asymmetries and legal restrictions).” Without such a model, it is difficult to see how a researcher might devise an identification strategy to detect collusion.

Despite the challenges, the insight of [Yinger’s \(1981\)](#) model offers a starting point to examine competitive nature of this industry. That is, if a market is perfectly competitive, then commission rates should vary inversely with home price and positively with the brokerage cost. Suppose two houses have equal marketing costs. In this case, competitive pressures should cause brokerage fees to also be equal. More generally, fees should approximate costs in a competitive equilibrium. This will lead to a tapering of the commission rate for higher-value houses. One should expect that commission rates would vary drastically across geographic regions as housing supply and demand vary dramatically across markets. One should also expect that commission rates would change over time as the diffusion of the Internet significantly reduced the cost of providing brokerage services. Using the CEX, [Hsieh and Moretti \(2003\)](#) and [Han and Hong \(2011\)](#) found a lack of correlation between commission rate and house prices, indicating that commission rates may not be market-driven. In addition, the time variation in commission rates seems to be rather limited, suggesting that commission rates are not very responsive to the changes in technology either. On the other hand, using the MLS transaction data covering housing transactions in Baton Rouge, Louisiana, between 1985 and 1992, [Sirmans and Turnbull \(1997\)](#) found that the commission rates actually exhibit a surprising amount of variation in response to changes in market conditions, consistent with what a simple competitive pricing model predicts.

A first attempt to formally test collusion among real estate brokers is provided in [Levitt and Syverson \(2008b\)](#). They formalized the intuition about brokers’ collusive behavior in a dynamic collusion model where a traditional agent chooses whether to cooperate with a discount agent in a pending transaction. If he cooperates, he will earn the commission income from the sale, but face lower expected future commissions. The model implies that a larger discount factor, greater reductions in future commissions, and a smaller loss from current sales if agents choose not to cooperate, would all lead to a higher likelihood of cooperation, making it easier to sustain collusion. Using data from three local markets, they found that houses listed using flat-fee agents have longer expected time-on-market than observably similar houses sold by full-commission agents but ultimately sell for similar prices. These results are consistent with the hypothesis that traditional agents steer clients away from flat-fee-listed homes, hence providing the first evidence for collusive behavior among real estate agents.

13.8.3 Social inefficiency

The rigid percentage commission structure in the real estate brokerage industry has profound implications not only for the competitive nature of the market but also for the efficiency of resource allocation. The central idea in the efficiency literature is that if commission rates are in fact fixed (or inflexible) and product differentiation is limited, then low entry barriers would result in socially wasteful expenditures of resources by agents in cities with higher house prices. In particular, higher commission incomes in higher-priced areas are simply dissipated through the entry of real estate agents who compete with existing agents for a fixed number of house sales. In this section, we review the theoretical and empirical work related to this argument.

13.8.3.1 Theoretical work

The idea that too many resources are devoted to brokers' search activities goes back at least to [Yinger \(1981\)](#). The key issue he addressed is whether the manner in which brokerage services have traditionally been provided and priced is consistent with efficiency. The main finding is that the traditional absence of price competition among brokers may have led to an inefficiently large commitment of resources to the marketing and promotion, supported by excessive rates for consumers of brokerage services.

Motivated by Yinger's insight, [Crockett \(1982\)](#) extended this analysis from the agent level to the brokerage firm level. He showed that a brokerage firm can increase its profits by employing more agents as more agents are likely to convert more listings into transactions. On the other hand, agents' compensation is completely contingent on their transactions. Thus, a brokerage firm is willing to hire too many agents in an effort to capture a larger share of available listings. Crockett concluded that "when price competition is forestalled, competition among firms takes nonprice routes whose ultimate effect may be to promote inefficiency in the provision of brokerage services."

One thing omitted in Crockett's analysis is free entry. [Miceli \(1992\)](#) went a step further by incorporating entry of brokers. This allows him to simultaneously examine the effects of nonprice competition and free entry on the allocation of resources to brokerage. By comparing the housing market equilibrium with and without brokers, he found that brokers can create a welfare gain to buyers and sellers if the commission revenue is less than the extra search costs that buyers and sellers would have to spend if they transacted on their own. However, with free entry and the absence of price competition, individual brokers compete with one another on a nonprice basis to acquire a large share of available listings. As a result, they fail to minimize the average costs of transactions.

13.8.3.2 Empirical work

The empirical literature on entry and efficiency is a recent one. In an inspiring work, [Hsieh and Moretti \(2003\)](#) used the 5% sample of the Census of Population and Housing

in 1980 and 1990. They showed that when the average price of land in a city increases, (1) the fraction of real estate brokers in a city increases, (2) the productivity of an average real estate agent, measured by houses sold per agent and houses sold per hour, falls, and (3) the real wage of a typical real estate agent remains unchanged. This evidence provides strong support for the idea that entry is socially excessive in the absence of price competition. A legitimate concern for this interpretation is that the higher commissions in high housing cost cities may reflect the possibility that a broker has to spend more time matching buyers and sellers in such cities. For example, expensive houses may have thinner markets, making homebuyers spend more time searching before making a decision. Therefore, the correlation between housing prices and the productivity of realtors may reflect differences in the quality of the service provided by realtors. Hsieh and Moretti addressed this concern by showing that as the average price of housing in a city increases, there is only a small increase in the amount of time a buyer spends searching for a house, and the average time a house stays on the market falls. This helps rule out the concern about differences in service quality.

Motivated by Hsieh and Moretti's work, [Han and Hong \(2011\)](#) and [Barwick and Pathak \(2014\)](#) provide quantitative evidence of the extent of inefficiency in this industry. The empirical challenge to such an exercise is that relevant data on brokerage cost is unavailable, making it difficult for researchers to quantify welfare loss associated with resource misallocation in the brokerage industry. [Han and Hong \(2011\)](#) and [Barwick and Pathak \(2014\)](#) offer a solution based on the recent methodology developments in the structural IO literature (e.g., [Berry and Reiss, 2007](#)). In particular, since the observed entry decision is an indicator of the underlying profitability, one can in principle recover the cost estimates using the information on individuals' entry decisions as well as agents. However, these two papers differ in their measures of cost inefficiency. In [Han and Hong \(2011\)](#), inefficiency is measured by an excessive increase in the brokerage cost in order to compete with other agents for existing listings. Such cost includes paid advertisements, writing blogs, updating websites, and informal networking with potential clients. To the extent that the benefits resulted from these brokerage activities do not offset the committed resources, this type of nonprice competition is considered inefficient. In [Barwick and Pathak \(2014\)](#), inefficiency is measured by the amount of income entrants could have alternatively earned had they not worked as agents. This foregone income is an inefficiency cost since agents' entry mostly dilutes the business of existing agents without increasing the total output of the brokerage industry.

More specifically, [Han and Hong \(2011\)](#) estimate a rational expectation equilibrium, which is represented by fixed points in entry probabilities, where agents' beliefs about other agents' entry coincide with the entry choice made by each agent. They then estimate an equilibrium model by employing a nested pseudo-likelihood algorithm ([Aguirregabiria and Mira, 2002, 2007](#)). Using the 5% sample of the 2000 Census of Population and Housing, they find strong evidence for cost inefficiency under free entry, particularly attributable to wasteful nonprice competition. They perform counterfactual

experiments to investigate the welfare impact of antirebate rules that have often been criticized for discouraging price competition. They find that rebate bans are welfare-reducing, not only because they suppress price competition from discount brokers but also because they encourage excessive entry by full-commission brokers. In an average metropolitan area with antirebate policies, removing these rebate bans would reduce real estate agents' revenues, thereby decreasing the equilibrium number of agents by 5.14% and reducing total brokerage costs by 8.87%.

[Barwick and Pathak \(2014\)](#) advance this literature further by specifying and estimating a richer structural model of dynamic entry and exit decisions for real estate agents. Using a comprehensive data set of agents and transactions from 1998 to 2007 in the Greater Boston area, they find that there is a strong business-stealing effect in that entry does not increase sales probabilities or reduce the time it takes for properties to sell, but rather decreases the market share of experienced agents. Motivated by these empirical patterns, they build on upon the existing dynamic discrete choice literature (e.g., [Aguirregabiria and Nevo, 2010](#)) and estimate an elegant dynamic entry and exit model. This allows them to identify the amount of income entrants could have earned had they not worked as agents based on the information about entry and exit decisions of agents combined with their observed commission revenue. The estimates imply that agents' forgone income is about 80% of their observed revenue. Using these estimates, they further compute counterfactual results from a series of interesting policy experiments. A one-half reduction in the commission rate leads to a 73% increase in the number of houses each agent sells and benefits consumers by about \$2 billion. House price appreciation in the first half of the 2000s accounts for 24% of overall entry and a 31% decline in the number of houses sold by each agent. Low-cost programs that provide information about past agent performance have the potential to increase overall productivity and generate significant social savings.

As discussed earlier, the main challenge for the empirical work on social inefficiency is that costs associated with brokerage are generally not reported in any public data source. [Hsieh and Moretti \(2003\)](#) deal with this in their analysis by looking at declines in average broker productivity, as measured by houses sold per agent. The two structural papers take the alternative approach of backing out costs from structural entry models. This illustrates the potential usefulness of taking a structural approach to these issues. However, the heterogeneity of housing units and the heterogeneity of brokers themselves have not been addressed simultaneously in a structural context yet because of computational difficulties. With heterogeneous brokers matched to particular properties, it is possible that a fraction of the resources devoted to attracting is not pure dissipation. This calls for future research.

13.9. INCENTIVE ISSUES IN REAL ESTATE BROKERAGE

The previous section concluded with a discussion of a particular sort of inefficiency of the brokerage industry: the dissipative use of resources by brokers. This section will focus on

a different sort of inefficiency, one associated with incentive conflicts between brokers and the sellers and buyers of houses. Consider a principal–agent setting where sellers and buyers are principals and their respective brokers are agents. The essential feature of the principal–agent problem is the presence of asymmetrical information. There are two types of asymmetrical information in the real estate brokerage market: first, a client cannot observe how much effort his agent is putting into selling his property (referred to as *hidden action or moral hazard*); and second, a client does not know how knowledgeable the agent is concerning the state of the market or how skillful the agent is (referred to as *hidden type or adverse selection*). The literature has mostly focused on incentive issues resulting from the first type of information asymmetry. For example, the percentage commission structure gives the broker only a small portion of the marginal benefits from additional effort. It thus fails to align the incentives of the broker with the interests of the client. In this section, we will review both theoretical and empirical work related to incentive issues in the residential real estate brokerage industry.

13.9.1 Incentive misalignment in real estate brokerage: Theory

Broadly speaking, the real estate brokerage literature is informed by an important and substantial literature on the distortion of incentives (e.g., Gruber and Owings, 1996; Hubbard, 1998; Garmaise and Moskowitz, 2004; Mehran and Stulz, 2007). More specifically, an extensive and growing literature examines the consequence of brokers' incentive misalignment, with a particular focus on the moral hazard issues between sellers and their agents. Within this context, the central issue for the seller is to design a commission contract that aligns her own interest with the interest of her broker. A typical brokerage contract in North American markets involves a fixed commission rate and a fixed duration. Specifically, if a broker sells the property during the period of the contract, she receives a fixed percentage of the price at the time of sale; otherwise, she receives nothing. A long line of research has shown that the typical contract features serious incentive problems.

Yavaş (1996) showed that in a standard setting with one principal and one agent, a net listing contract is the only commission structure that would elicit the efficient intensity level from the agent. In Yavaş's model, with a net listing contract, the broker is the residual claimant. Any percentage commission structure where the broker receives less than 100% of the residual will fail to offer adequate incentives for optimal search effort from brokers.³⁶ This would lead brokers to leave their own homes on the market longer and

³⁶ A related idea is that contracts would be more efficient if they specified a commission rate that increases with house price. For instance, instead of paying 6% of the sales price of a \$500,000 house, or \$30,000 in total, a contract offering 30% of the excess of price over \$400,000 would give the broker stronger incentives. However, brokerage is motivated by the lack of information of house sellers, so it is not clear how sellers might decide which sort of contract they might want to offer. In any case, such contracts do not appear to exist in the market.

sell at a higher price, compared to homes they sell for their clients (e.g., [Rutherford et al., 2005](#); [Levitt and Syverson, 2008a](#)). The inability of percentage contracts to provide sufficient agent incentives is noted by [Zorn and Larsen \(1986\)](#), [Anglin and Arnott \(1991\)](#), [Yavaş \(1995\)](#), and [Rutherford et al. \(2004\)](#), among others.³⁷

The conclusion that the percentage commission contract produces agency problem relies on an assumption that one broker works for one seller. In a one-seller-one-broker setting, the broker always allocates time between leisure and selling the asset of the single client. Since the marginal value of leisure does not depend on the commission, a percentage commission less than 100% induces the broker to consume excessive leisure and thereby to spend insufficient effort.

However, once we depart from the one-seller-one-broker assumption and allow competition among brokers, it is possible that the percentage commission scheme will no longer produce agency problems. In this section, we will first review the theoretical literature that incorporates broker competition and then review the empirical literature on a series of incentive misalignment problems caused by the current commission structure.

Two notable papers on broker competition are [Williams \(1998\)](#) and [Fisher and Yavaş \(2010\)](#). Both papers offer a model of search under a percentage commission structure and show that the compensation structure generates no agency problem. The difference between the two papers is that the results of the former paper depend crucially on the competition among agents for new listings, while the latter paper requires multiple agents to compete to sell any of the listings available on the market.

[Williams \(1998\)](#) introduced a model of competitive equilibrium for brokers where there is no agency problem between brokers and their current clients. The innovations in his model are the following assumptions: (1) multiple brokers, (2) possibly multiple sellers per broker, (3) costly search for both buyers and new sellers, and (4) a competitive equilibrium among brokers. His model can be summarized as follows: in each period, the representative broker chooses between labor and leisure. To each client, the broker then allocates the same time or effort and, finally, any remaining time he spends searching for new clients. For each client, the broker selects the same reservation price, independent of the number of current clients. The somewhat surprising result is that each broker spends the same time or effort selling each client's asset and selects the same reservation price as he would for his or her own assets. In other words, the standard contract with a fixed-percentage commission produces no agency problems in equilibrium between a broker and current clients. This is due to the fact that competition among brokers equates each agent's marginal productivity of effort spent with each listing to the marginal productivity of searching for new listings and the two marginal productivities are proportional to the commission rate, implying that the optimal effort spent for each client is independent of

³⁷ See also [Larsen and Park \(1989\)](#) for an empirical analysis of nonuniform commissions.

the commission rate. However, this equilibrium outcome is not efficient because time spent searching for new clients has a private value to each agent but no value to their clients or to other agents. In addition, since the average arrival rate of buyers at each house depend on the average time allocated by all brokers to other houses, this creates a further deviation between the equilibrium outcome and the Pareto optimal outcome.

Unlike Williams (1998), Fisher and Yavaş (2010) considered a setting where the percentage commission system produces no agency problem even in the absence of the competition for new listings. Their innovation is to assume multiple agents compete to sell any of the listings with the MLS. The first agent to procure a buyer receives the entire commission, while other agents obtain zero commission. Under this type of compensation rule, an agent's search intensity not only impacts the seller's payoff but also other agents' payoffs. In particular, the race among agents induces each agent to spend too much search effort to sell a given home in order to improve the probability that she is the procuring agent (as in Mortensen, 1982). On the other hand, a commission rate less than 100% would induce the broker to spend insufficient effort selling a home, as predicted by standard principal-agent analysis. In equilibrium, these two inefficiencies offset each other and result in efficient effort levels. While Yavaş and Fisher's model rationalizes the optimality of the percentage commission contract, it does not explain the observed uniform commission rates across markets and over time, because the efficient level of commission rates in their model varies with house price, size of the market, and brokerage costs, as they note.

The models in Williams (1998) and Fisher and Yavaş (2010) are important, both because competition among brokers is a constant feature of the real estate brokerage market and because the agency problem is a first-order issue that concerns buyers and sellers in this market. It is also worth noting that both models require strong assumptions. For example, both papers assume that brokers and houses are identical. With heterogeneous brokers competing for selling a particular house, such search effort is no longer purely dissipative. In addition, Fisher and Yavaş (2010) assume that the seller lists his property with the MLS system without acquiring a listing agent first. This allows them to eliminate the search by agents for new listings, which is a crucial assumption in Williams' model. While there are some signs that sellers are slowly gaining opportunities to post their listings directly on the MLS, this innovative feature is still not widely used in the current North American markets. Once we deviate from these assumptions, the standard agency problem may reemerge.

In much of the literature, broker effort generates better offers in an unspecified way. Ehrlich (2013) is an exception. He presents a model of one-sided seller search where the seller is uncertain about the state of the market, while the broker is not. When an offer arrives, the broker advises the seller on whether to take it or not. Ehrlich obtains the interesting result that a seller should always believe a broker who counsels patience, but not necessarily a broker who counsels acceptance. This captures the advisory role

of real estate brokers and the moral hazard arising from the percentage rent contract. [Stacey \(2013\)](#) is another exception, considering intermediation by brokers in a contracting framework. His results show how agents can improve liquidity, which is consistent with some of the empirical work reviewed here.

13.9.2 Empirical work on incentive misalignment

An extensive and growing empirical literature examines the consequence of the misalignment between goals of real estate brokers and those of home sellers. [Figure 13.3](#), discussed in [Section 13.1](#), illustrates many of the aspects of intermediation that we will consider in this section. The earlier work on this topic has focused on the effects of full-commission versus split-commission agreements ([Munneke and Yavaş, 2001](#)) and exclusive-agency (EA) agreement versus exclusive-right-to-sell (ERTS) agreement ([Rutherford et al., 2001](#)). More recent work has examined the effects on selling price and time-on-market of agent-owned versus client-owned properties ([Rutherford et al., 2005](#); [Levitt and Syverson, 2008a](#)), MLS-listed versus FSBO properties ([Hendel et al., 2009](#)), and properties sold by traditional agents versus discounted agents ([Levitt and Syverson, 2008b](#); [Bernheim and Meer, 2013](#)). One common thread running through this line of research is that the current commission arrangements have resulted in a distortion of brokers' incentives, which in turn affects how much a house is sold for and how long it takes to sell.

13.9.2.1 Broker-owned versus client-owned properties

The key incentive issue is whether brokers, as motivated by their commissions, will undertake the efficient amount of effort on behalf of their clients. While the current percentage commission scheme gives brokers an incentive to work on behalf of sellers to obtain high prices, it does not provide sufficient incentive for brokers to achieve the highest price possible, as brokers only get a small percentage of the marginal benefits from the additional effort they put in. This argument has been well established in [Anglin and Arnott \(1991\)](#) and [Geltner et al. \(1991\)](#), among others. As [Levitt and Syverson \(2008a\)](#) emphasized, a broker “has strong incentives to sell the house quickly, even at a substantially lower price.”

This hypothesis, while intuitively appealing, can be empirically difficult to test, because brokers' effort is hard to measure. A clever approach designed by [Rutherford et al. \(2005\)](#) and [Levitt and Syverson \(2008a\)](#) is to compare brokers' performance when they sell their own houses and when they sell their clients' houses. [Rutherford et al. \(2005\)](#) used a sample of condominiums—a housing market segment that presumably has less degree of heterogeneity. They found that brokers received a premium of 3–7% when selling their own condominiums in comparison with similar condominiums owned by their clients. In a similar spirit, [Levitt and Syverson \(2008a,b\)](#) used a larger data set of single-family home sales where a portion of the sample was broker-owned to

measure for agency problems. They found that agent-owned homes sold for a price premium of about 4–5% over client-owned homes and stay on the market longer. They also used a Herfindahl index to measure the degree of heterogeneity of houses in a given city block. They found that the highest premium was in the most heterogeneous blocks.

Various other papers have taken similar approaches. [Huang and Rutherford \(2007\)](#) examined the effect on the sales price and the time-on-market for Realtor versus non-Realtor listed homes, where Realtor refers to members of the NAR. They found that houses sold by realtors on the MLS sell for more and sell faster than those sold by agents without that designation. [Bian et al. \(2013\)](#) considered a related issue, the impact of the number of houses that a broker is selling on price and liquidity. They showed that price is lower and that properties take longer to sell when there are more other listings competing for a broker's time. [Jia and Pathak \(2014\)](#) investigated whether sellers who pay higher commissions experience different sales outcomes. They found that a higher commission is associated with a higher likelihood of sale, a modest impact on time-on-market, and overall no effect on the sales price, consistent with high commission agents selling properties at lower prices to increase the likelihood of selling. Together, these studies provide useful evidence for the conflicts of incentives in the real estate brokerage market resulting from the percentage commission arrangement.

13.9.2.2 Broker-listed versus FSBO properties

An alternative way to examine the importance of commission incentives is to compare the sales performance of properties sold by brokers and properties sold by owners themselves. In recent years, with the diffusion of the Internet, the practice of FSBO sales (“for sale by owner”—referring to homes that are sold without using a listing broker) has gained increasing popularity. Thus, in principle, researchers can quantify the benefits of hiring a broker by examining price and liquidity effects of broker-listed versus FSBO properties. The estimated premium can then be compared with the commission fees, permitting an evaluation of the efficiency of the commission contract.

Although the task for estimating the price premium of the brokers' services relative to the FSBO sounds straightforward, in recent years, a debate has ensued over how to transform the estimated price premium into a measure of the net benefit of brokerage services. For example, [NAR Report \(2005\)](#) found that FSBO houses sold for a median price of \$198,200 and those sold through a broker went for a median price of \$230,000. The report concludes that using an agent brings sellers a significantly higher price (16%). A legitimate concern with such analysis is that it is not clear whether sellers and their houses are otherwise similar for broker-listed and FSBO homes; hence, measured differences in sales price may reflect a combination of effects.

Alternatively, one could estimate the price effects of realtor service relative to the FSBO sales, controlling for house characteristics and market conditions. However, this

strategy itself cannot convincingly control for the fact that the use of a broker is highly correlated with the characteristics of homes and homeowners. In most markets, FSBO sellers constitute a small, highly selected group with potential unusual characteristics and inclinations. According to [Evans \(2003\)](#), during the first quarter of 2004, 44% of all FSBO homes were never placed on the open market, as the buyer and seller knew each other in advance. In addition, FSBO sellers tend to be older and less wealthy ([National Association of Realtors, 2002](#)). Some prior studies employ sample selection corrections, but identification is driven entirely by functional form assumptions rather than exclusion restrictions. Also, some earlier studies employed data sets that were extremely small and somewhat peculiar.

[Hendel et al. \(2009\)](#) compared sales of MLS-listed homes sold through traditional full-service brokers to sales of homes listed on an FSBO website in Madison, Wisconsin, in 1998–2004. Their analysis is noteworthy because their data set is reasonably large, contains many FSBO transactions, and spans a 7-year period, which allows them to control for both homes and household fixed effects. Controlling for differences in-house and seller characteristics, they found that listing on the MLS does not yield a price premium relative to listing on [FSBOMadison.com](#). However, listing on the MLS with a traditional broker does shorten the time it takes to sell a house and is more likely to ultimately result in a transaction. They also found evidence of endogenous sorting and reported that impatient sellers are more likely to list with the high commission, high service option. Note that the data set they used comes from one unique market where the FSBO sales account for about 25% of the total sales. One cannot generalize their conclusion beyond this market, as the penetration rates of the FSBO and the driving forces behind the diffusion of FSBO vary widely across markets and over time.

13.9.2.3 Traditional brokers versus discounted brokers

Between the two extremes of using a full-service broker and selling a house with no broker assistance whatsoever, there exist various intermediate options. An important example of this is the use of a discounted brokerage service, such as one that places clients' homes in the MLS for a fixed fee. Given the great degree of heterogeneity in buyer and seller demands for broker services, the practice of unbundling MLS listings with other real estate broker services can potentially improve consumer welfare. For this reason, it has gained support from both the Federal Trade Commission and the Department of Justice. This leads naturally to the question of how much value the brokerage services provide to sellers, beyond the value from being listed on the MLS. The answer to this is not obvious. On one hand, brokers offer useful knowledge and expertise. On the other hand, brokers may not represent the seller's interests given the principal-agent problem that we described earlier.

[Bernheim and Meer \(2013\)](#) shed light on this debate by studying Stanford Housing Listings. The houses in their sample are not listed on MLS. They compared the

performance of listings that are assisted by brokers versus FSBO sales. They found that sellers realize similar prices but sell less quickly when they elect not to hire a broker. As noted above in a different context, [Levitt and Syverson \(2008b\)](#) compared the performance between full-commission real estate brokers and discount brokers. They found that time-on-market is longer for houses sold with the assistance of less costly brokers, but sales prices are not significantly different. The common finding that the use of a broker has no impact on sales price seems to suggest that the incentive misalignment problem is sufficiently severe that it offsets the positive benefits brought by brokers' information advantage.

13.9.2.4 Exclusive agency versus exclusive right to sell brokers

So far, we have discussed an important agency problem that originates from the percentage commission structure. Going one step further, the literature has shown that the level of effort that brokers spend on their clients' houses also depends on the nature of the listing contract, the relationship between listing and cooperating agents, and the commission sharing arrangement between brokers and their affiliated brokerage firms. We now begin discussing these factors.

The agency relationship between the broker and the seller is formalized in the listing contract. According to [Rutherford et al. \(2001\)](#), the exclusive right to sell (ERTS) listing is the most common listing arrangement. With this listing, the broker receives a commission if the house sells or if a signed purchase-and-sale agreement is obtained prior to the expiration of the listing contract. This type of listing contracts is generally preferred by brokers because it offers the strongest guarantee of a commission. Another listing arrangement, exclusive agency (EA), allows the seller to avoid a commission if the seller is responsible for achieving the sale. However, similar to the ERTS listing, the listing broker receives the commission if any real estate broker achieves the sale prior to contract expiration.

Using a search-theoretic framework, [Rutherford et al. \(2001\)](#) showed that both contract types would yield the same price, while the EA contract would generate a greater effort level both from the broker and from the seller, hence resulting in faster sales than the ERTS contract. Using MLS data between 1994 and 1997 from Dallas–Fort Worth, Texas, they found that houses sell more quickly and at marginally lower prices under EA contracts than for ERTS contracts, consistent with what the theory suggested. Using the same data set, [Rutherford et al. \(2004\)](#) further looked at differences between housing market segments defined by price ranges. They found that for lower-priced houses, properties stay on the market for longer time and sell at a discount if the sale is done by brokers. On the other hand, sellers of higher-priced houses are better served by the EA contracts.

A related issue is the duration of the real estate listing contract. In North America, a typical contract specifies a fixed period of exclusivity. [Anglin and Arnott \(1991\)](#) noted that there is variation in the length of the period, with the most common durations falling

between 61 and 90 days. Contract duration has a number of effects. [Anglin and Arnott \(1991\)](#) showed that contracts that vary in duration and commission rate can be used to separate agents of different ability, with the low-ability agents more willing to sacrifice commission for longer duration. In [Miceli's \(1989\)](#) analysis of the optimal duration of brokerage contracts, a shorter contract may better align incentives. The idea is that agents will put forth greater effort in order to complete a sale prior to the expiration of the contract. A further issue, one not considered in the literature, is that a shorter duration may fail to give the broker proper incentives to carry out relationship-specific investments in the marketing of a particular house. See [Joskow \(1987\)](#) for an analysis of this in a different context. On the empirical side, [Waller et al. \(2010\)](#) found that time-on-market is longer when the listing contract is longer, consistent with Miceli's model.

13.9.2.5 Broker representation

When the listing broker receives the commission income from the seller, under the MLS guidelines, he then (equally) splits the commission with the cooperating broker who works on the buyer's side. This is because historically, the cooperating broker was viewed as a subagent of the seller and represented the seller's best interest. So it was logical for the seller to compensate the subagent. The obvious incentive problem with such an arrangement is that, while it gives brokers an incentive to work on behalf of the seller to obtain a higher price, it also creates perverse incentives for the broker to work on behalf of the buyer (see [Lindeman, 2004](#)).

The advent of buyer brokerage in many states has changed the legal representation in this relationship. In this case, the broker working with the buyer is no longer a subagent of the seller, but rather an agent of the buyer—referred to as a buyer broker. Using 1996 NAR data, [Elder et al. \(2000\)](#) found that buyer brokers appear to reduce search time but have no effect on price for buyers. Buyers with higher opportunity search costs and who are less knowledgeable about local market conditions are more likely to seek buyer brokers. [Curran and Schrag \(2000\)](#) also looked at the effect of buyer brokerage, showing that buyer brokerage lowers buyers' search cost and improves buyers' negotiation position in the case of high-end properties. However, as [Yavaş and Colwell \(1999\)](#) pointed out, as long as buyer brokers are paid a percentage of sales price, their interests are not aligned with those of the buyers.

Compared to the sellers, buyers are often considered more active in the search and matching process. This is because they can make decisions on the extensive margins, such as whether to search or not, whether to bid or not, and whether to buy or not. Recognizing the differences between buyers and sellers, [Miceli \(1991\)](#) proposed that sellers pay a fixed fee to a broker to list a property on the MLS and then a commission to only the broker who locates a buyer. Based on this arrangement, [Yavaş and Colwell \(1999\)](#) further proposed that the seller may hire a broker separately to assist with negotiations and that the buyer may separately hire a buyer broker for property showings and to assist in

negotiation. However, as [Miceli et al. \(2000\)](#) pointed out, such an arrangement still has not addressed the agency problems associated with buyer brokers being compensated based on the sales price.

The incentive issues present in the agency relationship are particularly severe when buyers and sellers are represented by the same brokerage office—so-called in-house transactions. In-house transactions account for about 20% of home transactions in a typical North American housing market ([Han and Hong, 2014](#)). In theory, in-house transactions could create information efficiency and reduce transaction costs, leading to an efficient match between homebuyers and sellers. However, it is also possible that agents may promote in-house transactions for their own financial interest. In particular, since matching internal listings with internal buyers helps clear inventories faster and increase the chance of securing both ends of a transaction, brokerage firms often pay a higher commission to reward agents engaged in in-house transactions. Quite naturally, these in-house transactions reflect agents' strategic efforts, creating a distortion in the home transaction process that benefits agents rather than homebuyers and sellers.

Strategic in-house transactions, if present, have at least two deleterious effects on homebuyers and sellers. First, in the search stage, real estate agents may misguide buyers (sellers) by directing their interest to internal listings (buyers), resulting in a suboptimal choice for consumers. Second, in the negotiation stage, an apparent conflict of interest arises from having the same agency represent both buyers and sellers, making it impossible for the agency to help one party without hurting the other. For these reasons, many jurisdictions have now introduced disclosure requirements for dual agency in order to help consumers avoid undisclosed and unintended dual agency relationship. The legislation requires brokerages and agents to inform both buyers and sellers about the nature of dual agency relationships in writing.

The effect of dual agency on the negotiation stage has been examined by a number of studies. For example, [Gardiner et al. \(2007\)](#) found that dual agency reduced the sales price and the time-on-market and that both effects were weaker after a law change in Hawaii in 1984 that required full disclosure of dual agency. Using repeated sales properties, [Evans and Kolbe \(2005\)](#) showed little influence by the presence of dual agent on property price. Similarly, using 10,888 transactions in Long Island, New York, in 2004–2007, [Kadiyali et al. \(2014\)](#) find that dual agency has an overall null effect on sales price. More recently, [Johnson et al. \(2014\)](#) employ MLS transaction data from Johnson County, Indiana, for the period June 1, 2000, through May 31, 2010, and find that dual agency has a null effect on sale price. By further controlling for the ownership of the property, they also find that dual agency is associated with a price premium on agent-owned properties but a price discount on government- and bank-owned properties.

Like the existing literature on the real estate brokerage, the aforementioned studies use sales price and time-on-market as key measures for assessing brokers' performance. While these measures reflect sellers' interest, they do not reflect the quality of match

between homebuyers and the houses they purchase, and hence cannot address the effect of dual agency on the search stage of the home transaction process. [Han and Hong \(2014\)](#) address this question by developing a structural model of in-house transactions and recover the match values that a homebuyer could obtain from internal listings and from external listings. Doing so allows them to evaluate the economic harm that the incentive misalignment brings to homebuyers. In a world where agents' interests are fully aligned with homebuyers' interests, there should be no efficiency loss associated with in-house transactions since all transactions represent the best matching outcome for buyers. On the other hand, if agents strategically interfere with the allocation of houses to individuals, buyers' benefits would be inevitably sacrificed, and a suboptimal match would be generated.

Using the home transaction data from a large North American metropolitan area, [Han and Hong \(2014\)](#) find that agents are more likely to promote internal listings when they are financially rewarded and that such effect is weaker after the implementation of a legislation that requires agents to disclose their dual agency relationship to their clients in writing. In particular, about 64.3% of in-house transactions are explained by efficient matching, while the remaining are likely due to agents' strategic promotion—the latter causes significant utility loss for homebuyers. They also find that the legislation has weakened the impact of agents' strategic promotion on the home matching process, which accounts for 70% of the decrease in in-house transactions after the regulatory change.

13.9.2.6 Full-commission brokers versus split-commission brokers

Once the brokers receive commissions, they split commissions with the affiliated brokerage firms. Some brokerage firms give their brokers a predetermined ratio of their commission revenue, referred to as split-commission agents. Others, such as RE/MAX, allow their brokers to retain 100% of their commission income and require a fixed amount of upfront fees instead, referred to as full-commission agents. Because a full-commission agent receives a larger commission from selling any given listing than a split-commission agent, the former would attract more listings and spend more effort on each listing. However, as the listings increase, an agent's marginal productivity declines. Using a search-theoretic framework, [Munneke and Yavaş \(2001\)](#) showed that in equilibrium, there will be no difference in either property price or selling time between full- and split-commission agents. Using home sales data in the Athens, Georgia, area, [Munneke and Yavaş \(2001\)](#) found that RE/MAX agents obtain significantly more listings than other agents. Furthermore, there is no statistically significant difference between two types of agents with respect to the time it takes to sell a listing and the sales price. Using a different data set, [Allen et al. \(2003\)](#) found that residential properties marketed by full-commission agents are sold more quickly and at a premium relative to properties sold by split-commission agents.

In both studies, RE/MAX agents are used to represent the full-commission agents (and hence “more productive” agents), while non-RE/MAX agents are used to represent split-commission agents (and hence “less productive agents”). [Salter et al. \(2007\)](#) applied the Markowitz portfolio optimization theory to determine the optimal combination of full (less risky) and split agents (more risky). They showed that firms need to retain over 10% of full-commission agents. This gives a practical explanation for the diversification of agents within a firm, and it illustrates the synergies that full and split agents create when working for the same firm. It also implies that a simple categorization of agents based on the RE/MAX affiliation may create significant measurement error and specification problems. Based upon an actual determination of each agent’s specific compensation arrangement via a survey of qualifying brokers, [Johnson et al. \(2008\)](#) reexamined the relationship between the commission split structure and agent performance using sales data from the Montgomery, Alabama, market area. They found that full-commission agents sell their listed properties faster and at premiums compared to split-commission agents.

A limitation of this line of research is that the causal direction of the relation between commission incentives and agent performance is unclear. Rather than higher commission incentives producing better performance, it may be the case that more productive agents select to be full-commission agents ([Munneke and Yavaş, 2001](#)). For example, [Zumpano et al. \(2009\)](#) found that younger, male, more experienced, more past income agents tend to prefer to be full-commission agents. It would be worthwhile for researchers to analyze this important question using an equilibrium approach or select appropriate instrumental variables to address the selection issues.

13.9.2.7 Incentive issues in rental markets

We have thus far focused on intermediaries in housing purchases. There is also intermediation in the rental market, but it takes a rather different form. The vast majority of North American brokerage agreements for house sales create an exclusive relationship between the broker and seller. In leasing, there are both exclusive and nonexclusive contracts. In addition, for sales, the seller pays brokerage fees. For leasing, sometimes the landlord pays, and sometimes, the new tenant pays.

[Bar-Isaac and Gavazza \(2014\)](#) consider the determination of contractual form for leasing agreements using data from Manhattan. They find a pattern of contracts that is consistent with incentive issues being important in leasing brokerage. More atypical units (as defined by [Haurin, 1988](#)) are more likely to be sold under exclusive contracts. In thinner markets, it is more difficult to find well-matched tenants, and an exclusive contract gives stronger incentives to the broker. The landlord paying the fees for rent-stabilized apartments allows the charging of a high initial rent, which is important when the growth of rents is restricted.

13.9.2.8 Conclusion

If there is one theme that unifies the large body of empirical work on incentive issues, it is that researchers have focused on estimating the treatment effects of various forms of intermediation on housing transaction outcomes such as price and liquidity. However, the intermediation forms (e.g., for sale by owner vs. represented by a broker) are not randomly assigned to houses and their sellers. The variation in the data, thus, is not experimental in its nature. There is good reason to believe that sellers who adopt unusual intermediation strategies may be different in ways that impact price and marketing time. Much attention has been paid and more should be given in the future to the issue of arriving at estimates that have causal interpretations.

13.9.3 Mitigating the incentive problems

Despite the existence of a large literature on incentive misalignment in real estate, there is relatively little research on how these incentive problems might be mitigated. [Williams \(1998\)](#) showed that competition among agents for new listings can correct incentives, in some situations yielding first-best solutions. [Fisher and Yavaş \(2010\)](#) showed that a different sort of competition—the competition among agents to sell an existing listing—also may give a first-best solution. Essentially, the race among agents encourages too much effort, which counterbalances the tendency under percentage commissions to put forth too little effort.

Improvements in information will also tend to address incentive problems. Some of this is a competition effect. The Internet has allowed the creation and expansion of new business models for real estate intermediation. For instance, most of FSBO sales rely on the Internet, and they compete directly with sales intermediated by real estate brokers. As above, competition will help to correct incentive problems. Furthermore, agency problems in general stem from asymmetric information. Improvements in information technology, such as Trulia and Zillow, reduce asymmetries and thus presumably improve efficiency. There is not yet precise empirical work that examines this issue.

One very natural response by brokers themselves to the observation that there may be incentive problems is that brokers have reputations and bad behavior would be punished. [Shi and Tapia \(2014\)](#) considered this issue. They did this by comparing outcomes of sales between a group of sellers who leaves the area and a group who does not. The former is clearly less likely to offer referrals. They found that this group experiences faster sales and lower price.

Another approach to this issue would be to devise a contract with better incentive properties. The traditional 6% commission is a rather low-powered incentive contract. The alternative approach practiced in Vancouver of a 7% commission on the first \$100,000 of house price and a lower commission on the rest makes the incentives even lower in power. [Geltner et al. \(1991\)](#) discussed issues of incentives and commented on the puzzling absence of stronger incentive contracts.

13.10. CONCLUSIONS

This chapter has reviewed a very large body of research that has considered how housing markets operate. This has involved reviewing the literature on housing search, bargaining, auctions, and intermediation. One conclusion that is worth emphasizing is that housing markets are illiquid. They clear through both price and various measures of liquidity, including seller and buyer time-on-market and the amount of search effort put forth by buyers and sellers.

Another clear conclusion that we can make is that housing markets are very different from other markets. Housing search, unlike labor or goods search, typically involves households acting as both buyers and sellers. This has important implications for the nature of equilibrium. Similarly, real estate brokers are different from other intermediaries in many ways. One is that they will frequently act in their own capacities as homebuyers and home sellers, something that we do not see for the intermediaries in other markets. This gives a way to learn about agency in a very fundamental way from considering brokerage. These are, of course, only a few of the ways that the specific characteristics of housing impact how housing markets operate.

For future theoretical research, there is a strong case for considering models with robust microfoundations. Most of the other issues considered in this chapter involve in some way imperfect or incomplete information, and the modeling of this is central to the theoretical analysis. Understanding, for instance, the role of the asking price requires explicit treatment of the information that it conveys to potential searchers. There is also a strong case for working with equilibrium models when possible. In models surveyed here, we have seen how equilibration has important implications for liquidity, market dynamics, and various other significant phenomena. It is worth noting that in many of these equilibrium models, a stationary framework has been the dominant theoretical paradigm. However, a nonstationary equilibrium model would be particularly useful for assessing the dynamics of the housing market, such as the overshooting phenomenon in the short run and the stickiness of asking price. Of course, any theoretical model, whether general or partial equilibrium, stationary or nonstationary, must be empirically grounded in order for the analysis to capture the unique institutions and features of the housing market.

Theory is also important for empirical work because it provides guidance for the empirical identification of a causal relationship. Housing is well known to be heterogeneous, as are households and brokers. A successful identification strategy must resolve econometric issues that arise from selection and sorting based on unobserved heterogeneity of any of these types. The chapter has identified numerous areas where identifying the effects of the market and intermediation factors illustrated in [Figures 13.1–13.3](#) is complicated by this sort of endogeneity problem.

These problems will be difficult to resolve, but empirical progress will presumably be made easier by the availability of new data sources. Compared to traditional sources, new

data sets are sometimes much larger and are often panels. Furthermore, the new data sources sometimes contain information not only on transaction outcomes but also on the fine structure of the transaction process such as the bargaining process, bidding, and buyer search activities. The new data sources thus provide a much more comprehensive picture of the search, bargaining, and intermediation process among buyers, sellers, and brokers, allowing researchers to look at old puzzles from new perspectives and to address a variety of new issues.

ACKNOWLEDGMENTS

We thank the Social Sciences and Humanities Research Council of Canada and the University of Toronto Centre for Real Estate for financial support. We also thank Gilles Duranton, Vernon Henderson, Paul Carrillo, Stuart Rosenthal, Joseph Williams, and participants of the Toronto Handbook of Regional and Urban Economics Conference for helpful comments. We further thank Yousuf Haque and Jessica Burley for helpful work as research assistants.

REFERENCES

- Adams, P.D., Kluger, B.D., Wyatt, S.B., 1992. Integrating auction and search markets: the slow Dutch auction. *J. Real Estate Financ. Econ.* 5 (3), 239–253.
- Aguirregabiria, V., Mira, P., 2002. Swapping the nested fixed point algorithm: a class of estimators for discrete Markov decision models. *Econometrica* 70, 1519–1543.
- Aguirregabiria, V., Mira, P., 2007. Sequential estimation of dynamic discrete games. *Econometrica* 75, 1–53.
- Aguirregabiria, V., Nevo, A., 2010. Recent developments in empirical IO: Dynamic demand and dynamic games, Working paper.
- Albrecht, J., Anderson, A., Smith, E., Vroman, S., 2007. Opportunistic matching in the housing market. *Int. Econ. Rev.* 48 (2), 641–664.
- Albrecht, J., Gautier, P., Vroman, S., 2012. Directed search in the housing market. Working paper.
- Allen, M.T., Faircloth, S., Forgey, F., Rutherford, R.C., 2003. Salespersons compensation and performance in the housing market. *J. Acad. Financ.* 1, 62–71.
- Anderson, R.I., Lewis, D., Springer, T.M., 2000a. Operating efficiencies in real estate: a critical review of the literature. *J. Real Estate Lit.* 8 (1), 1–18.
- Anderson, R.I., Lewis, D., Zuppano, L.V., 2000b. X-inefficiencies in the residential real estate market: a stochastic frontier approach. *J. Real Estate Res.* 20 (1), 93–104.
- Anderson, R.I., Lewis, D., Zuppano, L.V., 2000c. Residential real estate brokerage efficiency from a cost and profit perspective. *J. Real Estate Financ. Econ.* 20 (3), 295–310.
- Anenberg, E., Bayer, P., 2013. Endogenous Sources of Volatility in Housing Markets: The Joint Buyer-Seller Problem (No. w18980). National Bureau of Economic Research.
- Anglin, P.M., 1997. Determinants of buyer search in a housing market. *Real Estate Econ.* 25 (4), 567–589.
- Anglin, P.M., Arnott, R., 1991. Residential real estate brokerage as a principal-agent problem. *J. Real Estate Financ. Econ.* 4 (2), 99–125.
- Anglin, P., Arnott, R., 1999. Are brokers' commission rates on home sales too high? A conceptual analysis. *Real Estate Econ.* 27 (4), 719–749.
- Anglin, P., Rutherford, R., Springer, T., 2003. The trade-off between the selling price of residential properties and time-on-the-market: the impact of price setting. *J. Real Estate Financ. Econ.* 26 (1), 95–111.
- Arnold, M.A., 1999. Search, bargaining and optimal asking prices. *Real Estate Econ.* 27 (3), 453–481.
- Arnott, R., 1987. Economic theory and housing. In: Mills, E.S. (Ed.), *Handbook of Regional and Urban Economics*, vol. 2. North Holland, Amsterdam, pp. 959–988.

- Ashenfelter, O., Genesove, D., 1992. Testing for price anomalies in real estate markets. *Am. Econ. Rev.* 82, 501–505.
- Bar-Isaac, H., Gavazza, A., 2014. Brokers' contractual arrangements in the Manhattan residential rental market. Working paper.
- Bartlett, R., 1981. Property rights and the pricing of real estate brokerage. *J. Ind. Econ.* 30, 79–94.
- Barwick, P., Pathak, P., 2014. The impact of commissions on home sales in Greater Boston. *Am. Econ. Rev. Pap. Proc.* 100, 475–479.
- Battu, H., Ma, A., Phimister, E., 2008. Housing tenure, job mobility and unemployment in the UK. *Econ. J.* 118 (527), 311–328.
- Bayer, P., Geissler, C., Roberts, J., 2011. Speculators and middlemen: the role of flippers in the housing market. NBER Working paper Series, 16784.
- Benjamin, J.D., Jud, G.D., Sirmans, G.S., 2000. What do we know about real estate brokerage? *J. Real Estate Res.* 20, 5–30.
- Benjamin, J.D., Chinloy, P., Winkler, D.T., 2007. Sorting, franchising and real estate brokerage firms. *J. Real Estate Financ. Econ.* 34 (2), 189–206.
- Bernheim, B.D., Meer, J., 2013. Do real estate brokers add value when listing services are unbundled? *Econ. Inq.* 51 (2), 1166–1182.
- Berry, S., Reiss, P., 2007. Empirical models of entry and market structure. In: Armstrong, M., Porter, R.H. (Eds.), *Handbook of Industrial Organization*, vol. 3. Elsevier, Amsterdam, pp. 1845–1886.
- Bian, X., Turnbull, G., Waller, B., Wentland, S., 2013. How many listings are too many? The impact of agent inventory externalities on selling price and liquidity of client properties. Working paper.
- Braswell, M.K., Poe, S.L., 1992. The residential real estate brokerage industry: a proposal for reform. *Am. Business Law J.* 30 (2), 271–334.
- Bulow, J., Klemperer, P., 2009. Why do sellers (usually) prefer auctions? *Am. Econ. Rev.* 99 (4), 1544–1575.
- Burnside, C., Eichenbaum, M., Rebelo, S., 2011. Understanding booms and busts in housing markets. National Bureau of Economic Research Working paper 16734.
- Campbell, J.Y., Giglio, S., Pathak, P., 2011. Forced sales and house prices. *Am. Econ. Rev.* 101, 2108–2131.
- Caplin, A., Leahy, J., 2011. Trading frictions and house price dynamics. *J. Money Credit Bank.* 43 (2), 283–303.
- Caplin, A., Chan, S., Freeman, C., Tracy, J., 1997. *Housing Partnerships*. MIT Press, Cambridge.
- Carney, M., 1982. Costs and pricing of home brokerage services. *Real Estate Econ.* 10 (3), 331–354.
- Carrillo, P.E., 2008. Information and real estate transactions: the effects of pictures and virtual tours on home sales. George Washington University. Working paper.
- Carrillo, P.E., 2012. An empirical stationary equilibrium search model of the housing market. *Int. Econ. Rev.* 53 (1), 203–234.
- Carrillo, P.E., 2013. To sell or not to sell: measuring the heat of the housing market. *Real Estate Econ.* 41 (2), 310–346.
- Carrillo, P.E., de Wit, E.R., Larson, W., 2014. Can tightness in the housing market help predict subsequent home price appreciation? Evidence from the U.S. and the Netherlands. *Real Estate Econ.* Forthcoming.
- Carroll, W., 1989. Fixed-percentage commissions and moral hazard in residential real estate brokerage. *J. Real Estate Financ. Econ.* 2 (4), 349–365.
- Case, K., Shiller, R., 1989. The efficiency of the market for single-family homes. *Am. Econ. Rev.* 79 (1), 125–137.
- Case, K., Shiller, R., 1988. The behavior of home buyers in boom and post-boom markets. *N. Engl. Econ. Rev.* 11, 29–46.
- Case, K., Shiller, R., 2003. Is there a bubble in the housing market? *Brook. Pap. Econ. Act.* 2, 299–342.
- Chauvet, M., Gabriel, S., Lutz, C., 2014. Fear and loathing in the housing market: evidence from search query data. Working paper.
- Chen, Y., Rosenthal, R.W., 1996a. On the use of ceiling-price commitments by monopolists. *RAND J. Econ.* 27 (2), 207–220.
- Chen, Y., Rosenthal, R.W., 1996b. Asking prices as commitment devices. *Int. Econ. Rev.* 37, 129–155.
- Chow, Y.L., Hafalir, I.E., Yavas, A., 2014. Auction versus negotiated sale: evidence from real estate sales. *Real Estate Econ.*
- Coulson, N.E., Fisher, L.M., 2002. Tenure choice and labour market outcomes. *Hous. Stud.* 17 (1), 35–49.

- Coulson, N.E., Fisher, L.M., 2009. Housing tenure and labor market impacts: the search goes on. *J. Urban Econ.* 65 (3), 252–264.
- Courant, P.N., 1978. Racial prejudice in a search model of the urban housing market. *J. Urban Econ.* 5 (3), 329–345.
- Crockett, J.H., 1982. Competition and efficiency in transacting: the case of residential real estate brokerage. *Am. Real Estate Urban Econ. Assoc.* 10 (2), 209–227.
- Cubbin, J., 1974. Price, quality, and selling time in the housing market. *Appl. Econ.* 6 (3), 171–187.
- Curran, C., Schrag, J., 2000. Does it matter whom an agent serves? Evidence from recent changes in real estate agency law. *J. Law Econ.* 43, 265–284.
- Dachis, B., Duranton, G., Turner, M.A., 2011. The effects of land transfer taxes on real estate markets: evidence from a natural experiment in Toronto. *J. Econ. Geogr.* 11, 1–28.
- Diamond, P.A., 1982. Wage determination and efficiency in search equilibrium. *Rev. Econ. Stud.* 49, 217–227.
- Díaz, A., Jerez, B., 2013. House prices, sales, and time-on-market: a search-theoretic framework. *Int. Econ. Rev.* 54 (3), 837–872.
- Ehrlich, G., 2013. Price and time to sale dynamics in the housing market: the role of incomplete information. Working paper.
- Elder, H.W., Zumpano, L.V., Baryl, E.A., 1999. Buyer search intensity and the role of the residential real estate broker. *J. Real Estate Financ. Econ.* 18 (3), 351–368.
- Elder, H.W., Zumpano, L.V., Baryl, E.A., 2000. Buyer brokers: do they make a difference? Their influence on selling price and search duration. *Real Estate Econ.* 28 (2), 337–362.
- Evans, D.S., 2003. Antitrust economics of multi-sided platform markets. *Yale J. Regul.* 20, 325.
- Evans, R., Kolbe, P., 2005. Homeowners' repeat sales gains, dual agency and repeated use of the same agent. *J. Real Estate Res.* 27 (3), 267–292.
- Fisher, L.M., Yavaş, A., 2010. A case for percentage commission contracts: the impact of a “race” among agents. *J. Real Estate Financ. Econ.* 40 (1), 1–13.
- Fisher, J., Gatzlaff, D., Geltner, D., Haurin, D., 2003. Controlling for the impact of variable liquidity in commercial real estate price indices. *Real Estate Econ.* 31 (2), 269–303.
- Ford, J.S., Rutherford, R.C., Yavaş, A., 2005. The effects of the internet on marketing residential real estate. *J. Hous. Econ.* 14 (2), 92–108.
- FTC Report, 1983. The Residential Real Estate Brokerage Industry. Federal Trade Commission.
- FTC Report, 2007. Competition in the Real Estate Brokerage Industry. U.S. Department of Justice and Federal Trade Commission.
- Gardiner, J., Heisler, J., Kallberg, J.G., Liu, C.H., 2007. The impact of dual agency. *J. Real Estate Financ. Econ.* 35 (1), 39–55.
- Garmaise, M., Moskowitz, T., 2004. Confronting information asymmetries: evidence from real estate markets. *Rev. Financ. Stud.* 17 (2), 405–437.
- Geltner, D., Kluger, B., Miller, N.G., 1991. Optimal price and selling effort from perspectives of the broker and seller. *J. Am. Real Estate Urban Econ. Assoc.* 19 (1), 1–24.
- Genesove, D., Han, L., 2012a. Search and matching in the housing markets. *J. Urban Econ.* 72, 31–45.
- Genesove, D., Han, L., 2012b. Measuring the thinness of real estate markets. Working paper.
- Genesove, D., Mayer, C., 1997. Equity and time to sale in the real estate market. *Am. Econ. Rev.* 87 (3), 255–269.
- Genesove, D., Mayer, C., 2001. Loss aversion and seller behavior: evidence from housing markets. *Q. J. Econ.* 116 (4), 1233–1260.
- Glaeser, E.L., Nathanson, C.G., 2015. Housing bubbles. In: Duranton, G., Henderson, J.V., Strange, W.C. (Eds.), *Handbook of Urban and Regional Economics*, vol. 5. Elsevier, Amsterdam, pp. 699–751.
- Glaeser, E.L., Gyourko, J., Morales, J., Nathanson, C.G., 2014. Housing dynamics: an urban approach. *J. Urban Econ.* 81, 45–56.
- Glower, M., Haurin, D.R., Hendershott, P.H., 1998. Selling time and selling price: the influence of seller motivation. *Real Estate Econ.* 26 (4), 719–740.
- Goetzmann, W., Peng, L., 2006. Estimating house price indexes in the presence of seller reservation prices. *Rev. Econ. Stat.* 88 (1), 100–112.
- Goolsby, W.C., Childs, B.J., 1988. Brokerage firm competition in real estate commission rates. *J. Real Estate Res.* 3 (2), 79–85.

- Green, R.K., Vandell, K.D., 1998. Optimal asking price and bid acceptance strategies for residential sales. Working paper.
- Gruber, J., Owings, M., 1996. Physician financial incentives and cesarean section delivery. *RAND J. Econ.* 27 (1), 99–123.
- Guasch, J.L., Marshall, R.C., 1985. An analysis of vacancy patterns in the rental housing market. *J. Urban Econ.* 17 (2), 208–229.
- Han, L., 2010. The effects of price risk on housing demand: empirical evidence from US markets. *Rev. Financ. Stud.* 23 (11), 3889–3928.
- Han, L., Hong, S.H., 2011. Testing cost inefficiency under free entry in the real estate brokerage industry. *J. Bus. Econ. Stat.* 29 (4), 564–578.
- Han, L., Hong, S.H., 2014. In-House Transactions in the Real Estate Brokerage Industry. Rotman School of Management, University of Toronto, Working paper.
- Han, L., Strange, W., 2013. Bidding wars for houses. *Real Estate Econ.* 41 (3), 1–32.
- Han, L., Strange, W., 2014. What is the role of the asking price for a house? Working paper.
- Harding, J.P., Rosenthal, S.S., Sirmans, C., 2003. Estimating bargaining power in the market for existing homes. *Rev. Econ. Stat.* 85 (1), 178–188.
- Haurin, D., 1988. The duration of marketing time of residential housing. *Real Estate Econ.* 16 (4), 396–410.
- Haurin, D.R., Gill, H.L., 2002. The impact of transaction costs and the expected length of stay on homeownership. *J. Urban Econ.* 51 (3), 563–584.
- Haurin, D.R., Haurin, J.L., Nadauld, T., Sanders, A.B., 2010. List prices, sale prices, and marketing time: an application to U.S. housing markets. *Real Estate Econ.* 38 (4), 659–685.
- Head, A., Lloyd-Ellis, H., 2012. Housing liquidity, mobility, and the labour market. *Rev. Econ. Stud.* 79 (4), 1559–1589.
- Head, A., Lloyd-Ellis, H., Sun, H., 2014. Search and the dynamics of house prices and construction. *Am. Econ. Rev.* 104 (4), 1172–1210.
- Heckman, J.J., 1979. Sample selection bias as a specification error. *Econometrica* 47, 153–161.
- Hendel, I., Nevo, A., Ortalo-Magné, F., 2009. The relative performance of real estate marketing platforms: MLS versus FSBO Madison.com. *Am. Econ. Rev.* 99 (5), 1878–1898.
- Horowitz, J.L., 1992. The role of the list price in housing markets: theory and an econometric model. *J. Appl. Econ.* 7 (2), 115–129.
- Hosios, A.J., 1990. On the efficiency of matching and related models of search and unemployment. *Rev. Econ. Stud.* 57 (2), 279–298.
- Hsieh, C.T., Moretti, E., 2003. Can free entry be efficient? Fixed commissions and social waste in the real estate industry. *J. Polit. Econ.* 111 (5), 1076–1122.
- Huang, B., Rutherford, R., 2007. Who you going to call? Performance of realtors and non-realtors in a MLS setting. *J. Real Estate Financ. Econ.* 35 (1), 77–93.
- Hubbard, T., 1998. An empirical examination of moral hazard in the vehicle inspection market. *RAND J. Econ.* 29, 406–426.
- Jia, P., Pathak, P., 2014. The costs of free entry: an empirical study of real estate agents in Greater Boston. *RAND J. Econ.* Forthcoming.
- Johnson, K.H., Zumpano, L.V., Anderson, R.I., 2008. Intra-firm real estate brokerage compensation choices and agent performance. *J. Real Estate Res.* 30, 423–440.
- Johnson, K., Lin, Z., Xie, J., 2014. Dual agent distortions in real estate transactions. *Real Estate Econ.* Forthcoming.
- Joskow, P.L., 1987. Contract duration and relationship-specific investments: empirical evidence from coal markets. *Am. Econ. Rev.* 77, 168–185.
- Jovanovic, B., 1982. Truthful disclosure of information. *Bell J. Econ.* 13, 36–44.
- Jud, G.D., Frew, J., 1986. Real estate brokers, housing prices, and the demand for brokers. *Urban Stud.* 23, 21–31.
- Kadiyali, V., Prince, J., Simon, D.H., 2014. Is dual agency in real estate a cause for concern? *J. Real Estate Financ. Econ.* 48 (1), 164–195.
- Kim, S., 1992. Search, hedonic prices and housing demand. *Rev. Econ. Stat.* 74, 503–508.
- Knoll, M.S., 1988. Uncertainty, efficiency, and the brokerage industry. *J. Law Econ.* 31, 249–263.

- Krainer, J., 2001. A theory of liquidity in residential real estate markets. *J. Urban Econ.* 49, 32–53.
- Krainer, J., LeRoy, S.F., 2002. Equilibrium valuation of illiquid assets. *Econ. Theory* 19 (2), 223–242.
- Krishna, V., 2009. *Auction Theory*. Academic Press, New York.
- Kurlat, P., Stroebe, J., 2014. Testing for information asymmetries in real estate markets. Working paper.
- Landvoigt, T., Piazzesi, M., Schneider, M., 2013. The Housing Market(s) of San Diego.
- Larsen, J.E., Park, W.J., 1989. Non-uniform percentage brokerage commissions and real estate market performance. *Real Estate Econ.* 17 (4), 422–438.
- Lester, B., Visschers, L., Wolthoff, R., 2013. Competing with asking prices. Working paper.
- Levitt, S.D., Syverson, C., 2008a. Market distortions when agents are better informed: the value of information in real estate transactions. *Rev. Econ. Stat.* 90 (4), 599–611.
- Levitt, S., Syverson, C., 2008b. Antitrust implications of home seller outcomes when using flat-fee real estate agents. *Brookings-Wharton Papers on Urban Affairs*, 2008.
- Lindeman, B., 2004. Attrition of agency in real estate brokerage. *J. Am. Acad. Bus.* 4 (1/2), 377–383.
- Liu, C.H., Nowak, A., Rosenthal, S., 2014. Bubbles, post-crash dynamics, and the housing market. Working paper.
- Lusht, K.M., 1996. A comparison of prices brought by English auctions and private negotiations. *Real Estate Econ.* 24 (4), 517–530.
- Mayer, C.J., 1995. A model of negotiated sales applied to real estate auctions. *J. Urban Econ.* 38 (1), 1–22.
- Mayer, C.J., 1998. Assessing the performance of real estate auctions. *Real Estate Econ.* 26 (1), 41–66.
- McAfee, R.P., 1993. Mechanism design by competing sellers. *Econometrica* 61, 1281–1312.
- Mehran, H., Stulz, R., 2007. The economics of conflicts of interest in financial institutions. *J. Financ. Econ.* 85 (2), 267–296.
- Merlo, A., Ortalo-Magné, F., 2004. Bargaining over residential real estate: evidence from England. *J. Urban Econ.* 56, 192–216.
- Merlo, A., Ortalo-Magné, F., Rust, J., 2013. The home selling problem: theory and evidence. Working paper.
- Miceli, T.J., 1989. The optimal duration of real estate listing contracts. *Real Estate Econ.* 17 (3), 267–277.
- Miceli, T.J., 1991. The multiple listing service, commission splits, and broker effort. *Real Estate Econ.* 19 (4), 548–566.
- Miceli, T.J., 1992. The welfare effects of non-price competition among real estate brokers. *Real Estate Econ.* 20 (4), 519–532.
- Miceli, T.J., Pancak, K.A., Sirmans, C.F., 2000. Restructuring agency relationships in the real estate brokerage industry: an economic analysis. *J. Real Estate Res.* 20 (1/2), 31–47.
- Miceli, T.J., Pancak, K.A., Sirmans, C.F., 2007. Is the compensation model for real estate brokers obsolete? *J. Real Estate Financ. Econ.* 35 (1), 7–22.
- Miller, N.G., 1978. Time-on-market and selling price. *Real Estate Econ.* 6 (2), 164–174.
- Miller, N.G., Shedd, P.J., 1979. Do antitrust laws apply to the real estate brokerage industry? *Am. Business Law J.* 17 (3), 313–339.
- Miller, N.G., Sklarz, M.A., 1987. Residential property selling prices. *J. Real Estate Res.* 2 (1), 31–40.
- Moen, E.R., 1997. Competitive search equilibrium. *J. Polit. Econ.* 105 (2), 385–411.
- Mortensen, D.T., 1982. Property rights and efficiency in mating, racing, and related games. *Am. Econ. Rev.* 72 (5), 968–979.
- Mortensen, D.T., Pissarides, C.A., 1994. Job creation and job destruction in a theory of unemployment. *Rev. Econ. Stud.* 61, 397–415.
- Munch, J.R., Rosholm, M.M., Svarer, M., 2008. Home ownership, job duration, and wages. *J. Urban Econ.* 63 (1), 130–145.
- Munneke, H.J., Yavaş, A., 2001. Incentives and performance in real estate brokerage. *J. Real Estate Financ. Econ.* 22 (1), 5–21.
- National Association of Realtors, 2002. Annual Report. The National Association of Realtors, the research division.
- National Association of Realtors Report, 2005. Structure, Conduct, and Performance of the Real Estate Brokerage Industry. The National Association of Realtors, the research division.

- Ngai, L., Sheedy, K.D., 2014. Moving house. Working paper.
- Ngai, L., Tenreyro, S., 2014. Hot and cold seasons in the housing market. Working paper.
- Novy-Marx, R., 2009. Hot and cold markets. *Real Estate Econ.* 37 (1), 1–22.
- Ooi, J.T., Sirmans, C.F., Turnbull, G.K., 2006. Price formation under small numbers competition: evidence from land auctions in Singapore. *Real Estate Econ.* 34 (1), 51–76.
- Oswald, A., 1997. Theory of homes and jobs. Working paper.
- Owen, B.M., 1977. Kickbacks, specialization, price fixing, and efficiency in residential real estate markets. *Stan. Law Rev.* 29, 931–967.
- Peters, M., 1984. Bertrand equilibrium with capacity constraints and restricted mobility. *Econometrica* 52, 1117–1127.
- Peters, M., 1991. Ex Ante price offers in matching games: non-steady states. *Econometrica* 59, 1425–1454.
- Peters, M., Severinov, S., 1997. Competition among sellers who offer auctions instead of prices. *J. Econ. Theory* 75 (1), 141–179.
- Peterson, B.M., 2012. Fooled by search: housing prices, turnover, and bubbles: Bank of Canada. Working paper 2012–2013.
- Petrongolo, B., Pissarides, C.A., 2001. Looking into the black box: a survey of the matching function. *J. Econ. Lit.* 39, 390–431.
- Piazzesi, M., Schneider, M., 2009. Momentum traders in the housing market: survey evidence and a search model. *Am. Econ. Rev. Pap. Proc.* 99 (3), 406–411.
- Piazzesi, M., Schneider, M., Stroebe, J., 2013. Segmented housing search. Working paper. Stanford.
- Pissarides, C.A., 2000. *Equilibrium Unemployment Theory*, second ed. MIT Press, Cambridge, MA.
- Pissarides, C.A., 1985. Short-run equilibrium dynamics of unemployment vacancies, and real wages. *Am. Econ. Rev.* 75 (4), 676–690.
- Pryce, G., 2011. Bidding conventions and the degree of overpricing in the market for houses. *Urban Stud.* 48 (4), 765–791.
- Pryce, G., Gibb, K., 2006. Submarket dynamics of time to sale. *Real Estate Econ.* 34 (3), 377–415.
- Quan, D.C., 1994. Real estate auctions: a survey of theory and practice. *J. Real Estate Financ. Econ.* 9 (1), 23–49.
- Quan, D.C., 2002. Market mechanism choice and real estate disposition: search versus auction. *Real Estate Econ.* 30 (3), 365–384.
- Quigley, J.M., 1979. What have we learned about urban housing markets. In: Mieszkowski, P., Straszheim, M. (Eds.), *Current Issues in Urban Economics*. Johns Hopkins University Press, Baltimore, pp. 391–429.
- Rogerson, R., Shimer, R., 2011. Search in macroeconomic models of the labor market. In: Card, D., Ashenfelter, O. (Eds.), *Handbook of Labor Economics*, vol. 4. North Holland, Amsterdam, pp. 619–700.
- Rogerson, R., Shimer, R., Wright, R., 2005. Search-theoretic models of the labor market: a survey. *J. Econ. Lit.* 43 (4), 959–988.
- Rubinstein, A., 1982. Perfect equilibrium in a bargaining model. *Econometrica* 50 (1), 97–109.
- Rutherford, R.C., Springer, T.M., Yavaş, A., 2001. The impact of contract type of broker performance. *Real Estate Econ.* 29 (3), 389–409.
- Rutherford, R.C., Springer, T.M., Yavaş, A., 2004. The impact of contract type of broker performance: submarket effects. *Real Estate Econ.* 26 (3), 277–298.
- Rutherford, R.C., Springer, T.M., Yavaş, A., 2005. Conflicts between principals and agents: evidence from residential brokerage. *J. Financ. Econ.* 76 (3), 627–665.
- Salant, S.W., 1991. For sale by owner: when to use a broker and how to price the house. *J. Real Estate Financ. Econ.* 4 (2), 157–174.
- Salter, S., Johnson, K.H., Webb, J.R., 2007. Theory of the real estate brokerage firm: a portfolio approach. *J. Real Estate Portfolio Manag.* 13 (2), 129–138.
- Schnare, A.B., Kulick, R., 2009. Do real estate agents compete on price? Evidence from seven metropolitan areas. In: Glaeser, E.L., Quigley, J.M. (Eds.), *Housing Markets and the Economy: Risk, Regulation and Policy. Essays in Honor of Karl E. Case*. Lincoln Institute of Land Policy, Cambridge, MA, pp. 308–347.

- Schroeter, J.R., 1987. Competition and value-of-service pricing in the residential real estate brokerage market. *Quart. Rev. Econ. Business* 27, 29–40.
- Shi, S., 2001. Frictional assignment I: efficiency. *J. Econ. Theory* 98, 232–260.
- Shi, S., 2008. Search theory (new perspectives). In: Durlauf, S.N., Blume, L.E. (Eds.), *The New Palgrave Dictionary of Economics*, second ed. Palgrave Macmillan.
- Shi, L., Tapia, C., 2014. The discipline effects of concern for referrals: evidence from real estate agents. Working paper.
- Shiller, R., 1982. *Market Volatility*. MIT Press, Cambridge.
- Shiller, R., 1999. Human behavior and the efficiency of the financial system. In: Taylor, J.B., Woodford, M. (Eds.), *Handbook of Macroeconomics*, vol. 1. Elsevier, Amsterdam, pp. 1305–1340.
- Simon, H.A., 1955. A behavioral model of rational choice. *Q. J. Econ.* 69 (1), 99–118.
- Sirmans, C.F., Turnbull, G.K., 1997. Brokerage pricing under competition. *J. Urban Econ.* 41 (1), 102–117.
- Sirmans, C.F., Turnbull, G.K., Benjamin, J.D., 1991. The markets for housing and real estate broker services. *J. Hous. Econ.* 1 (3), 207–217.
- Springer, T.M., 1996. Single-family housing transactions: seller motivations, price, and marketing time. *J. Real Estate Financ. Econ.* 13 (3), 237–254.
- Stacey, D., 2013. Information, commitment, and separation in illiquid housing markets. Working paper.
- Stein, J.C., 1995. Prices and trading volume in the housing market: a model with down-payment effects. *Q. J. Econ.* 110 (2), 379–406.
- Stigler, G.J., 1961. The economics of information. *J. Polit. Econ.* 69 (3), 213–225.
- Stigler, G.J., 1962. Information in the labor market. *J. Polit. Econ.* 70 (5), 94–105.
- Stull, W.J., 1978. The landlord's dilemma: asking rent strategies in a heterogeneous housing market. *J. Urban Econ.* 5 (1), 101–115.
- Tracy, J., Schneider, H., 2001. Stocks in the household portfolio: a look back at the 1990s. *Curr. Issues Econ. Financ.* 7 (4), 1–6.
- Turnbull, G.K., Dombrow, J., 2007. Individual agents, firms, and the real estate brokerage process. *J. Real Estate Financ. Econ.* 35 (1), 57–76.
- Wachter, S.M., 1987. Residential real estate brokerage: rate uniformity and moral hazard. In: Jaffee, A. (Ed.), *Research in Law and Economics*, vol. 10. JAI Press, Greenwich, Connecticut, pp. 189–210.
- Waller, B.D., Brastow, R., Johnson, K.H., 2010. Listing contract length and time-on-market. *J. Real Estate Res.* 32 (3), 271–288.
- Wang, R., 1993. Auctions vs. posted price setting. *Am. Econ. Rev.* 83 (4), 838–851.
- Wang, R., 2011. Listing prices as signals of quality in markets with negotiation. *J. Ind. Econ.* 59 (2), 321–341.
- Wheaton, W.C., 1990. Vacancy, search, and prices in a housing market matching model. *J. Polit. Econ.* 98 (6), 1270–1292.
- White, L., 2006. The residential real estate brokerage industry: what would more vigorous competition look like? Working paper.
- Williams, J.T., 1998. Agency and brokerage of real assets in competitive equilibrium. *Rev. Financ. Stud.* 11 (2), 239–280.
- Williams, J.T., 2014. Housing markets with construction, screening, and focused search. Working paper.
- Wu, C., Colwell, P.F., 1986. Equilibrium of housing and real estate brokerage markets under uncertainty. *Real Estate Econ.* 14 (1), 1–23.
- Yavaş, A., 1992. A simple search and bargaining model of real estate markets. *Real Estate Econ.* 20 (4), 533–548.
- Yavaş, A., 1994. Middlemen in bilateral search markets. *J. Labor Econ.* 12 (3), 406–429.
- Yavaş, A., 1995. Seller-broker relationship as a double moral hazard problem. *J. Hous. Econ.* 4 (3), 244–263.
- Yavaş, A., 1996. Matching of buyers and sellers by brokers: a comparison of alternative commission structures. *Real Estate Econ.* 24 (1), 97–112.
- Yavaş, A., 2001. Impossibility of a competitive equilibrium in the real estate brokerage industry. *J. Real Estate Res.* 21 (3), 187–200.
- Yavaş, A., Colwell, P., 1999. Buyer brokerage: incentive and efficiency implications. *J. Real Estate Financ. Econ.* 18 (3), 259–277.

- Yavaş, A., Yang, S., 1995. The strategic role of listing price in marketing real estate: theory and evidence. *Real Estate Econ.* 23 (3), 347–368.
- Yinger, J., 1981. A search model of real estate broker behavior. *Am. Econ. Rev.* 71, 591–605.
- Zahirovic-Herbert, V., Turnbull, G.K., 2008. School quality, house prices and liquidity. *J. Real Estate Financ. Econ.* 37 (2), 113–130.
- Zietz, J., Sirmans, G.S., 2011. Real estate brokerage research in the new millennium. *J. Real Estate Lit.* 19 (1), 5–40.
- Zorn, T.S., Larsen, J.E., 1986. The incentive effects of flat-fee and percentage commissions for real estate brokers. *Real Estate Econ.* 14 (1), 24–47.
- Zuehlke, T.W., 1987. Duration dependence in the housing market. *Rev. Econ. Stat.* 69, 701–709.
- Zumpano, L.V., Johnson, K.H., Anderson, R.I., 2009. Determinants of real estate agent compensation Choice. *J. Hous. Econ.* 18 (2), 195–207.

CHAPTER 14

US Housing Policy

Edgar O. Olsen*, Jeffrey E. Zabel†

*Department of Economics, University of Virginia, Charlottesville, VA, USA

†Department of Economics, Tufts University, Medford, MA, USA

Contents

14.1. Introduction	888
14.2. Methods and Data	890
14.3. US Low-Income Rental Housing Policy	892
14.3.1 Justifications for low-income housing assistance	893
14.3.2 US housing programs	895
14.3.3 Three major social experiments	897
14.3.3.1 <i>Welfare to work voucher evaluation</i>	897
14.3.3.2 <i>MTO for fair housing demonstration program</i>	899
14.3.3.3 <i>Experimental housing allowance program</i>	903
14.3.4 Evidence on the performance of low-income housing programs	904
14.3.4.1 <i>Cost-effectiveness in providing housing</i>	904
14.3.4.2 <i>Housing consumption</i>	906
14.3.4.3 <i>Neighborhood occupied</i>	910
14.3.4.4 <i>Racial and economic integration</i>	912
14.3.4.5 <i>Consumption of other goods</i>	913
14.3.4.6 <i>Mean tenant benefit and subsidy</i>	914
14.3.4.7 <i>Distribution of benefits</i>	915
14.3.4.8 <i>Labor earnings and employment</i>	916
14.3.4.9 <i>Participation in other welfare programs</i>	923
14.3.4.10 <i>Children's educational outcomes</i>	923
14.3.4.11 <i>Recipient health</i>	925
14.3.4.12 <i>Household composition</i>	926
14.3.4.13 <i>Effects on neighbors of assisted households</i>	926
14.3.4.14 <i>Market prices</i>	932
14.3.4.15 <i>Crowding out of private provision</i>	933
14.3.4.16 <i>Miscellaneous</i>	934
14.3.5 Conclusion	935
14.4. US Homeownership Policy	938
14.4.1 Justifications for promoting homeownership	939
14.4.2 The homeownership rate in the United States	940
14.4.3 The history of the federal role in the mortgage market	943
14.4.3.1 <i>The evolving structure of mortgages</i>	947
14.4.3.2 <i>Low-income and minority homeownership policy</i>	951
14.4.3.3 <i>Securitization</i>	954

14.4.4	The impact of federal intervention in the mortgage market on homeownership rates	955
14.4.4.1	<i>Explaining the postwar surge in homeownership</i>	955
14.4.4.2	<i>The recent rise in homeownership</i>	958
14.4.4.3	<i>The impact of the affordable housing goals</i>	959
14.4.4.4	<i>The GSE impact on mortgage rates</i>	963
14.4.4.5	<i>The government role in the mortgage default and foreclosure process</i>	965
14.4.5	The government response to the housing crisis	971
14.4.6	Conclusion	975
14.5.	Conclusion	977
	References	978

Abstract

Governments throughout the world intervene heavily in housing markets, and most have multiple policies to pursue multiple goals. This chapter deals with two of the largest types of housing policies in the United States, namely, low-income rental assistance and policies to promote homeownership through interventions in mortgage markets. We describe the rationales for the policies, the nature of the largest programs involved, the empirical evidence on their effects, and the data and methods used to obtain them. Because the US government uses such a wide range of policies of these types, this evidence has lessons for housing policy in other countries.

Keywords

Housing policy, Low-income housing programs, Low-income housing assistance, Housing subsidies, Public housing, Housing vouchers, Homeownership rate, Mortgage markets, Foreclosure prevention, Promoting homeownership

JEL Classification Codes

H5, I3, R21, R28, R31, R38

14.1. INTRODUCTION

Governments throughout the world intervene heavily in housing markets. All pursue multiple goals, and most use multiple programs to pursue each goal.¹ Because there is much more evidence on the effects of these interventions in the United States, this chapter focuses on US housing policies. Specifically, it deals with two of the largest areas of government activity, namely, low-income rental assistance and policies to promote homeownership through interventions in mortgage markets. Government involvement in both areas began during the Great Depression, and the housing crisis that led to the recent recession resulted in many new programs in the latter category.

¹ The UN-Habitat's Adequate Housing Series, vols. 1–4 (<http://unhabitat.org/series/adequate-housing-series/>), describe housing policies throughout the world. Whitehead and Scanlon (2007) provided a more detailed account for European countries.

US governments also promote homeownership by many other means, most notably the preferential treatment of homeownership in the federal individual income tax, and they pursue other goals with policies such as fair housing laws, land use regulations, impact fees, rent control, and building and occupancy codes. [Fischel \(2000\)](#), [Glaeser and Gyourko \(2008\)](#), [Gyourko and Molloy \(2015\)](#), [Gyourko and Sinai \(2003\)](#), [Hilber and Turner \(2014\)](#), [Olsen \(1998\)](#), [Poterba and Sinai \(2011\)](#), [Rosen \(1979, 1985\)](#), [Schill \(2005\)](#), and [Yinger \(1995\)](#) provided good introductions to the evidence on these interventions.

Recent research has confirmed the primary deficiencies of the system of low-income housing assistance in the United States that have been known for decades, namely, its inequities and the substantial cost-ineffectiveness of project-based housing assistance that still dominates the system. However, the focus of research on low-income housing policies has changed markedly over the past 15 years. As a result, we have much better information on certain aspects of program performance such as the effect of the largest low-income housing programs on labor earnings and a wide range of effects of several important proposed reforms of the current system.

Nevertheless, important gaps in our knowledge remain. We have no high-quality evidence on the cost-effectiveness and most other outcomes of the largest new low-income housing programs, and we have no recent evidence on the performance of older programs that still account for a substantial minority of assisted households. For example, no attempt has been made in several decades to determine how much better or worse housing public housing tenants occupy than they would have occupied in the absence of housing assistance. Almost all evidence is based on data from the 1960s and 1970s, and the aging of the public housing stock raises serious doubts about the applicability of these results to the current situation.

A consistent message from the federal government since the 1920s is that homeownership is the “American Dream.” This has been the motivation behind many federal policies to promote homeownership. We will show in this chapter that, other than interventions during the significant economic crises defined as the Great Depression and Great Recession, state and federal policies aimed at increasing homeownership have generally been ineffective. In fact, one thing we have learned from the recent financial crisis is that pushing into homeownership households who should not be homeowners can lead to disastrous results; the American Dream can end up being the American Nightmare.

Analyses of the impact of state and federal interventions in the mortgage market have benefited greatly from access to increasingly larger and more detailed data on mortgage market activity including loan-level data on mortgage originations and performance. Furthermore, the latest research uses the recent advances in quasi-experimental methodology to identify exogenous variation in government policies to estimate of their causal impacts. We focus on these studies in our evaluation of the effectiveness of government policies in increasing the homeownership rate.

Section 14.2 discusses the advances in data and methods that have characterized the recent literature on evaluating US housing policies. Sections 14.2 and 14.3 cover low-income rental housing policies and government involvement in the mortgage market. In each case, we describe the justifications for the policies, the nature of the largest programs involved, the empirical evidence on their effects, and the data and methods used to obtain them. These areas differ markedly in the justifications for government action, the intended beneficiaries, the nature of the programs, and the outcomes studied.

14.2. METHODS AND DATA

Major advances in the literature on US housing policy have come through access to the vast data resources that have become available in recent years. Most studies that we survey use multiple data sources. For example, studies on the mortgage market often use large loan-level data sets that involve merging multiple data sources to link information on loan originations with loan-type, loan performance, and borrower characteristics. These samples often exceed 1 million observations.

The literature has also benefited from the application of a variety of excellent econometric methods.² It is well known that the gold standard for estimating the causal impact of policies on outcomes of interest is randomized controlled trials (RCTs). RCT involves a treatment (or policy) that is applied to a randomly drawn treatment group. Then, the comparison of outcomes between the treatment and the control groups can be attributed to the treatment or policy. In the 1970s, the Experimental Housing Allowance Program's (EHAP) Housing Assistance Demand Experiment conducted one of the oldest random assignment social experiments in order to study differences in the effects of alternative types of tenant-based rental housing assistance, and the recent Moving to Opportunity (MTO) for Fair Housing Demonstration Program and Welfare to Work Voucher evaluation have used this method to study a wide range of outcomes of rental housing policies. Some of the best other studies of these policies rely on data generated by random assignment for reasons of program administration rather than research.

Typically, though, there are no RCTs to evaluate the causal impacts of most government policies and programs. Fortunately, an important recent advance in empirical methodology in economics is the development and application of the quasi-experimental approach to causal inference (see Angrist and Pischke, 2009). This approach is based on the use of exogenous variation in key variables as a means for identifying causal impacts of treatments (or policies) on outcome variables of interest. Techniques that fall under the quasi-experimental umbrella include regression discontinuity (RD) design, the difference-in-difference framework, natural experiments, and matching techniques.

² See Baum-Snow and Ferreira (2015) for a more detailed review of using these methods to obtain causal inference in urban and regional economics.

RD is based on a treatment that is determined by the value of a predictor or forcing variable X being on either side of a threshold, c . Assume a unit receives the treatment if $X > c$. Then, units with values of X “near” the threshold c can be considered to be randomly assigned; those with values greater than c are in the treatment group, and those with values less than c are in the control group. This approach is used by a number of researchers to evaluate the affordable housing goals under the Community Reinvestment Act (CRA) and the Government-Sponsored Enterprises (GSEs) Act. These goals involve minimum percentage purchases or originations of mortgages in census tracts with median household income at or below a given threshold as compared to area median income (e.g., 80% or 90%). Then, it is possible to design an RD analysis based on census tracts just above or below this threshold.

The difference-in-difference framework uses within unit variation in explanatory variables to identify causal impacts on outcome variables. This is obtained by including unit fixed effects in the model. The framework mimics that of the RCT whereby the treatment effect is the difference in the difference in the outcome variable for the treatment and control groups before and after the treatment is administered. One way to identify the impact of a policy variable is to collect data from before and after the policy is implemented. This is often not possible with many housing policies that have been in existence for many years prior to the initial year of most housing data sets. If there is no change in the policy variable over time, then it is still possible to identify the policy impact using a border fixed effects approach. For example, when evaluating state-level policies, one can use as the unit of observation Metropolitan Statistical Areas (MSAs) that cross borders of states with different values of the policy variable. For example, this approach is used to identify the impact of the state-level law that requires judicial involvement in foreclosure proceedings.

Closely related to these procedures is the instrumental variables (IV) estimator that uses exogenous variation in the instrument to identify the causal impact of an endogenous regressor on the outcome variable. One source of instruments is natural experiments that arise when external circumstances result in something that is similar to a randomized experiment. This can occur when a policy or law is enacted over different periods of time, for different geographic areas, or applies to one group and not to another. For example, if a law is passed in one state and not in another and it can reasonably be assumed that one’s state of residence is a random event, then the law can be treated as a randomized experiment where the treatment/control group is the population in the state where the law was/was not passed.

One of the recent advances in IV is the interpretation of estimated impacts as local average treatment effects (LATEs). An instrument, Z , only captures a subset of the variation in the endogenous variable, X , that affects the outcome of interest; the variation in X covaries with Z , that is, the part of X that is related to Z . This is the subset of units or individuals that are incentivized by the treatment (or instrument). For example, when

using a census tract indicator of affordable housing goal eligibility just around the cutoff as an instrument for loan volume, the treatment effect only applies to census tracts that just meet the goal criteria (e.g., median household income no more than 90% of AMI). If there is heterogeneity in the treatment effect, it may not be possible to generalize from the LATE to the full population (in this case, all census tracts with median household income below 90% of AMI). This also applies to RD that uses a nonlinear change in the policy variable around a threshold value to identify the treatment effect. The key is that the change is only plausibly exogenous just around the threshold and so the treatment effect is only identified at the threshold value.

Another technique that is frequently used to obtain causal estimates is matching. This involves pairing treatment and control group members based on similar observable characteristics. Then, treatment effects are just the differences in mean outcomes for the matched treatment and control groups. Causal inference is predicated on treatment/control group assignment being independent of the outcome value. Given a large number of observable covariates, a common matching technique is the propensity score. This is a way of combining the covariates into a single index that makes it easier to match on.

Typically, housing policies are not exogenous, so these techniques are necessary to estimate the causal impacts of these policies. Many recent studies rely on one of these techniques as a means for identifying and estimating the causal impact of US housing policies on housing market and other outcomes. Our evaluation of the literature will put more weight on the studies that use one of these techniques as they are more likely to produce reliable results.

14.3. US LOW-INCOME RENTAL HOUSING POLICY

This section provides evidence on the performance of low-income rental housing programs in the United States. Evidence on the performance of US programs is much more abundant than evidence on the performance of programs in other countries, and most low-income housing assistance in the United States is tied to renting a unit. Because the US government uses such a wide range of methods to deliver housing assistance, this evidence has lessons for housing policy in other countries.

The most useful information for housing policy would be evidence on the differences in the effects of spending the same amount of money to assist the same people by different means. No study comes close to this ideal. Most studies focus on a small subset of outcomes of a single program. The few that compare programs focus on a few outcomes or don't account for differences in their cost. Putting together the information needed to design good housing policies from the existing literature resembles an archaeological dig. The dig reveals many gaps in our knowledge. Nevertheless, due to its importance for policy, we attempt to characterize the differences in the performance of the three broad types of low-income housing assistance in the United States—public housing,

privately owned subsidized projects, and housing vouchers. High-quality evidence on important issues available at the right time has arguably resulted in better policies ([National Research Council, 2008](#), chapter 9).

Because the desired outcomes of any program depend on its rationales, we begin with a discussion of the justifications for low-income housing assistance, and since their actual outcomes depend importantly upon the program's structure, we then provide a brief description of the programs involved.³ After that, we describe three major studies that have provided high-quality information on many outcomes of important housing programs or proposed housing policy reforms, namely, the Welfare to Work Voucher evaluation, the MTO for Fair Housing Demonstration Program, and the EHAP, and summarize their most important findings. Finally, we describe and assess the evidence on the main aspects of the performance of low-income housing programs that have been studied. These include their cost-effectiveness in providing housing, their effects on the types of housing and neighborhoods occupied by recipients, their consumption of other goods, the labor earnings and employment of adult recipients, participation in other welfare programs, educational outcomes for children and their later adult earnings, recipient health and household composition, the magnitude of recipient benefit relative to the subsidy provided, the effects of assisted households on their new neighbors, the effects of the programs on rents of unsubsidized units, and the extent to which the programs increase the number of housing units available.

The focus of research on low-income housing policy has changed markedly over the past 15 years. As a result, we have much better information on certain aspects of program performance such as the effect of the largest low-income housing programs on labor earnings and a wide range of effects of several important proposed reforms of the current system. However, important gaps in our knowledge remain. We have no high-quality evidence on the cost-effectiveness and most other outcomes of the largest new low-income housing programs, and we have no recent evidence on the performance of older programs that still account for a substantial minority of assisted households. For example, no attempt has been made in several decades to determine how much better or worse housing public housing tenants occupy than they would have occupied in the absence of housing assistance. Almost all evidence is based on data from the 1960s and 1970s, and the aging of the public housing stock raises serious doubts about the applicability of these results to the current situation.

14.3.1 Justifications for low-income housing assistance

Because recipients of assistance prefer an unrestricted cash grant to any in-kind subsidy with the same cost to donors, the justification for subsidizing housing consumption hinges on the preferences of others. The reasons for their preference might be altruistic

³ [Olsen \(2003\)](#) and [Weicher \(2012\)](#) provided more details.

or more narrowly self-interested. Unless nonrecipients prefer the outcomes of housing subsidies to the outcomes of cash grants, housing assistance cannot be justified.

The traditional argument for low-income housing assistance has been that people with higher incomes want to help low-income families and believe that the decision makers in some of these families undervalue housing for themselves or their children. This is an external effect that argues for providing assistance that induces its recipients to occupy better housing than they would choose if they were given equally costly unrestricted cash grants. The existence of minimum housing standards in all low-income housing programs reflects this sentiment. Another important argument is that people with higher incomes care about the children in these households and think that their parents devote too little of the family's resources to their children. Providing housing assistance rather than unrestricted cash grants directs more of the assistance to the children. To the best of our knowledge, the extent to which nonrecipients hold these views has not been studied.

It is sometimes argued that better housing confers tangible external benefits on people who don't live in it. There is little doubt that some external benefits of this type exist. For example, a house whose exterior is better kept confers a positive benefit on people who see it. However, the available evidence does not indicate substantial external effects of this sort, and hence, it seems unlikely that substantial public expenditures for low-income housing assistance can be justified based on these externalities alone, especially assistance that is not focused on the exteriors of houses and apartments. The subsection on the effects of low-income housing projects and families with housing vouchers on the neighbors of assisted households provides some evidence about tangible external effects.

The failure to think seriously about rationales for government action has led many to accept incoherent arguments for housing assistance (that is, a program that induces all recipients to consume more housing services than they would choose if given an equally costly unrestricted cash grant). For example, the most common argument for providing housing assistance among noneconomists in recent years has been that low-income households spend excessive fractions of their incomes on housing rather than that they live in inadequate housing. Some who offer this argument for housing assistance seem to believe that these households are forced to spend such a high fraction of their income on housing. They fail to realize how many vacant units of an appropriate size renting for less than the household's current expenditure are available in the same locality. They also fail to consider the possibility that these people would rather spend less on other goods than to live in worse housing or neighborhoods or at less convenient locations. Since the only negative consequence of spending a high fraction of income on housing is low consumption of other goods, some people who make this argument might be saying that these low-income households undervalue other goods relative to housing. If taxpayers want to help these households and feel this way about their choices, attaining an efficient allocation of resources that is preferred by everyone to the allocation in the absence of government action requires subsidies for all goods except housing to this group.

Although the preceding is an incoherent argument for housing assistance, it is not an incoherent argument for programs that are commonly called housing programs. Among families in the same economic circumstances, some spend small fractions of their income on housing and others spend large fractions. There is nothing inconsistent about believing that some people undervalue and others overvalue housing. These beliefs argue for a government program or combination of programs that induces some recipients to consume more housing services than they would choose if they were given equally costly cash grants and others to consume less housing services. Real housing programs in the United States change budget frontiers in ways that could have this effect. For example, some programs offer families an all-or-nothing choice of a particular dwelling unit at a below-market rent. This unit might be better or worse than the recipient would have chosen if offered an equally costly unrestricted cash grant. Therefore, real housing programs might provide housing subsidies to some recipients and nonhousing subsidies to others. Based on the rationale in this paragraph, such a program would be successful if it provided housing subsidies to people with the weakest taste for housing relative to other goods and nonhousing subsidies to people with the strongest taste for housing.

The rationales for government action are highly relevant for the design and assessment of government programs. Researchers who begin with a clear view about this matter are more likely to study issues that are important for developing good public policies.

14.3.2 US housing programs

The bulk of low-income housing assistance in the United States is funded by the federal government through a substantial number of programs that in total cost over \$50 billion a year.⁴ Unlike other major means-tested transfer programs in the United States, the system of low-income housing programs doesn't offer assistance to many of the poorest families who are eligible for them. Eligible families that want assistance must get on a waiting list when it is open for new applicants. Each local public housing authority and private entity that operates a subsidized project establishes its own system for deciding the order in which families on the waiting list are offered assistance within broad federal guidelines.

Most low-income housing assistance in the United States is for renting a unit, and all rental programs have minimum housing standards.⁵ The most important distinction between rental housing programs is whether the subsidy is attached to the dwelling unit (project-based assistance) or the assisted household (tenant-based assistance). If the

⁴ Olsen (2003, pp. 370–376) provided a short history of the development of the system.

⁵ The US government also provides housing assistance to low-income homeowners. Olsen (2007a) documented the bias of the current system of low-income housing assistance toward renting, provided a brief description of the major homeownership programs, and summarized the evidence on one aspect of their performance. Olsen and Ludwig (2013, pp. 218–221) assembled additional information about the performance of two of the largest homeownership programs. However, little is known about most effects of any of these programs.

subsidy is attached to a rental dwelling unit, each family must accept the particular unit offered in order to receive assistance and loses its subsidy if it moves to another unit unless it is able to obtain alternative housing assistance before moving. Each family offered tenant-based assistance is free to occupy any unit that meets the program's minimum housing standards, rents for less than the program's applicable ceiling (if any), and is affordable with the help of the subsidy and whose owner is willing to participate in the program. The family retains its subsidy if it moves to another unit meeting these conditions. Olsen (2003, pp. 399–404) described how these programs change the budget frontiers of families offered assistance.

The US Department of Housing and Urban Development (HUD)'s housing choice voucher program is the only significant program of tenant-based assistance. It is the second-largest low-income program, serving about 2 million households and accounting for about 30% of all households that receive low-income rental assistance.

There are two broad types of project-based rental assistance, namely, public housing and privately owned subsidized projects. Both types have usually involved new construction. In almost all other cases, they have required substantial rehabilitation of existing buildings. Many of these programs no longer subsidize the construction of new projects, but most projects built under them still house low-income households with the help of subsidies for their operation and renovation. Overall, project-based assistance accounts for about 70% of all households that receive low-income rental assistance.

Public housing projects are developed and operated by local public housing authorities established by local governments, albeit with substantial federal subsidies and regulations that restrict their choices. In the public housing program, government employees make most of the decisions made by unsubsidized for-profit firms in the private market—what to build, how to maintain it, and when to tear it down. Decisions about where to build projects have been heavily influenced by local political bodies. The public housing stock has declined by about 300,000 units since its peak in 1991. About 1 million households live in public housing projects.

Government agencies also contract with private parties to provide housing in subsidized projects. Most are for-profit firms, but not-for-profits have a significant presence. This is the largest part of the system, involving a number of different programs. The largest are the Low-Income Housing Tax Credit (LIHTC) funded through the federal tax system, HUD's Section 8 New Construction and Substantial Rehabilitation Programs and Section 236 Rental and Cooperative Housing Program, and US Department of Agriculture's Section 515/521 Program. Under these programs, in exchange for certain subsidies, private parties agree to provide rental housing meeting certain standards at restricted rents to eligible households for a specified number of years. The subsidy formulas are different for the different programs and lead to differences across programs in the input mix used to provide housing services. None of these programs provide subsidies to all suppliers who would like to participate. This is highly relevant for their

performance. In general, subsidies to selected sellers of a good have very different effects than subsidies to all sellers. About 4 million households live in projects of this type.

The subsidy received by the tenant is not necessarily the same as the subsidy paid to his or her landlord. From the perspective of the tenant, the subsidy is the difference between the market rent of the unit occupied and the tenant's rent.⁶ Occupants of subsidized housing projects are offered a particular unit with a particular market rent. Voucher recipients have some choice concerning the market rent of the unit occupied, and this affects their subsidy over a certain range of choices. Most recipients of housing assistance in the United States contribute 30% of their adjusted income toward the cost of their housing, and their subsidy is reduced by 30 cents for each additional dollar earned. The major exception is families living in newly built LIHTC projects. These families pay a rent that does not depend on their income and is always at least 30% of their income. Some tax credits are used to rehabilitate existing subsidized housing projects, and most occupants of these projects continue to pay 30% of their adjusted income in rent.

14.3.3 Three major social experiments

Three major social experiments have produced enormous amounts of evidence relevant for low-income housing policy. The Welfare to Work Voucher evaluation provides high-quality evidence on an unusually wide range of effects of HUD's largest low-income housing program compared with no housing assistance. Evidence from this experiment will be cited throughout the chapter. MTO provides reliable estimates of a similarly wide range of effects of an important potential reform of public housing, HUD's oldest low-income housing program. The older and much larger EHAP studied important effects of fundamental reforms of the current system, most importantly, implementing an entitlement housing voucher program. In real terms, this is the largest social experiment in the history of the US government. We describe and summarize the main results of MTO and EHAP in this section.

14.3.3.1 Welfare to work voucher evaluation

The Welfare to Work Voucher evaluation presents results from a random assignment experiment designed to estimate the effects of HUD's housing voucher program on a wide range of outcomes—adult earnings, employment rates, other employment outcomes, receipt of other public assistance, household size and composition, housing conditions, neighborhood characteristics, travel time to work, food expenditure, adult physical and mental health, and children's school performance, educational progress, behavioral problems, time use, delinquency, and risky behavior. In all, more than 100 outcomes were

⁶ The benefit to the tenant cannot exceed this amount, and the taxpayer cost always exceeds the subsidy to the tenant by at least the cost of administering the program.

studied, and impacts for many subgroups are reported (Abt Associates Inc. et al., 2006).⁷ Results are presented for periods up to 4.5 years after random assignment.

The experiment collected voluminous information about the 8731 eligible families who agreed to participate. These families were on the housing voucher waiting lists of housing authorities in Atlanta, Augusta (Georgia), Fresno, Houston, Los Angeles, and Spokane who had received aid from the Temporary Assistance for Needy Families (TANF) program or were eligible for it. TANF serves the poorest families with children, and almost half of voucher recipients nationally participate in this program. The data were collected from baseline and follow-up surveys, the administrative records of multiple programs, address tracking methods, public use data from the Census Bureau and Bureau of Labor Statistics, and qualitative interviews.

Randomly chosen families were assigned to a treatment group that would be offered a housing voucher immediately. Additional vouchers were allocated to the housing authorities for this purpose. Other families that agreed to participate prior to random selection formed the control group. Members of the control group were not without housing assistance over the course of the experiment. All remained on the housing voucher waiting list initially, and over the course of the experiment, about 41% of the households in the control group had used housing vouchers compared with 67% in the treatment group. At the time of the follow-up survey, about 37% of the households in the control group still had housing vouchers compared with 51% in the treatment group. Others in the control group lived in a subsidized housing project at baseline or moved into one before the follow-up survey. Therefore, the study's intent-to-treat (ITT) estimates are estimates of the average impact on the target population of the immediate offer of a housing voucher compared with a continuation of the current system of low-income housing assistance.

The immediate offer of a housing voucher to all families on the voucher waiting list would be a significant reform of the housing voucher program, and the voucher evaluation provides considerable information needed to assess it. A comprehensive assessment would require information on the extra cost incurred by taxpayers to provide housing vouchers sooner rather than later. It is important to realize that this reform falls well short of creating an entitlement housing voucher program for all eligible families. Because housing voucher waiting lists are not open for new applicants most of the time, they do not contain many eligible families who would use vouchers. Furthermore, the experiment was limited to families that had received aid from the TANF program or were eligible for it.

Generally speaking, the ITT estimates about 4 years after enrollment in the experiment indicated small positive effects in most respects. The negative effects were also

⁷ In 2011, the authors discovered that the subgroup impacts for two baseline housing status groups were incorrectly reversed in the appendix tables and text discussion of them. The electronic version of the report on HUD's Web site contains the correct results.

small, and most effects were not statistically significant. The primary exceptions were substantial reductions in the fraction of families homeless during the past year (33% reduction) and living with a friend or relative (31% reduction). About 14% fewer lived in crowded conditions and the treatment group's rent was 10% lower on average. The reason for the small ITT effects is that by the time of data collection for most outcomes, the difference in the nature of the housing assistance received by the control and treatment groups had narrowed considerably. The ITT effects were almost surely greater in the earlier years of the experiment.

The analysts developed a methodology for estimating the treatment-on-treated (TOT) effects for voucher recipients in the treatment group who would not have received vouchers with a continuation of the current system by the time of the follow-up survey. About 26% of all families are in this group. The methodology does not require strong assumptions because the differences in average outcomes between the treatment and the control groups are entirely due to this subgroup.

The reported TOT estimates are usually interpreted as estimates of the average effects of receipt of a housing voucher compared with no housing assistance. Strictly speaking, this is not correct even for the specified subgroup. Some control families in this subgroup surely lived in subsidized projects at the time of final data collection. About 13% of all families in the control group lived in such projects at baseline and some surely continued to live there at the time of the follow-up survey. Others moved to subsidized projects prior to final data collection. However, because it seems likely that relatively few control families in the specified group received housing assistance at the time of the follow-up survey and there is no better evidence on the effect of housing vouchers compared with no housing assistance, we have interpreted the study's TOT estimates in the usual manner in this chapter.

14.3.3.2 MTO for fair housing demonstration program

The primary purpose of MTO was to learn about the effects of living in a better neighborhood rather than the effects of a specific reform of low-income housing policy. However, the experiment was ideally designed to estimate the effects of two important potential reforms, and this chapter will focus on the results of the reform that has been studied with MTO data.

Like the voucher evaluation, MTO collected data on a wide range of outcomes. It has already spawned a substantial volume of high-quality research, and its rich data will undoubtedly be used for many more studies.⁸ MTO is very unusual in its duration. It collected data in Baltimore, Boston, Chicago, Los Angeles, and New York beginning

⁸ Sanbonmatsu et al. (2011, pp. 16–18) summarized the MTO research prior to the final evaluation with references to the leading studies. Kling et al. (2007) provided an authoritative analysis of the data collected for interim evaluation.

in 1994 and continuing through 2010. The experiment involved offering housing vouchers to occupants of public or private subsidized projects located in central city census tracts with a poverty rate greater than 40%.⁹ Specifically, eligible participants were randomly assigned to one of three groups: (1) those who were offered a Section 8 housing voucher on the condition that they occupy a unit in a census tract with a poverty rate of less than 10% for at least a year and adhere to the other voucher program requirements, (2) those who were offered regular Section 8 vouchers, and (3) those who initially remained in their current housing project but were entitled to pursue other forms of housing assistance. Households in the first group received substantial counseling and assistance in finding a private unit (costing about \$4500 per voucher used in 2014 prices). The experiment was not designed to estimate the effect of this aspect of the treatment separately from other aspects, but several studies have found that search assistance substantially increased voucher usage (Shroder, 2002a; Galiani et al., 2012).

The results of the experiment do not show the effect of a voucher offer or using a voucher compared with continuing to live in the initial housing project. Instead, they show their effects compared with a continuation of the current system. Members of the control and treatment groups were free to pursue other types of housing assistance and were sometimes forced to do it. By the end of the experiment, most members of the control group had left their public housing units. Indeed, about 42% of public housing units in MTO housing projects were demolished prior to data collection for the Final Impacts Evaluation. Their occupants were offered the option of a housing voucher or a vacant unit in another public housing project.

Table 14.1 indicates the type of housing assistance received by members of the control group and the regular voucher group at the times of the main data collection for the interim and final evaluations. Clearly, the experiment led to substantial differences in the type of housing assistance received, but by the time of final data collection, about a fourth of the members of the control group had housing vouchers and many in the treatment group no longer received them. About 62% of the families in the regular voucher treatment group used the vouchers offered initially and others in it got them later through regularly channels, but by the time of the final evaluation, less than 45% still used them.

MTO produced an interim report based on outcomes 4–7 years after families entered the program and a final evaluation based on outcomes 10–15 years after entry (Orr et al., 2003; Sanbonmatsu et al., 2011). Each has great value. The results in the last year of MTO are particularly valuable for the adult outcomes of children who were very young in the early years of the experiment. These are more important than early outcomes for children such as test scores whose importance is what they foretell about future outcomes. The

⁹ In 2000, only 11% of poor people and 3% of all people in the United States lived in census tracts with poverty rates this high. However, in the mid-1990s, 36% of public housing tenants lived in such census tracts (Newman and Schnare, 1997, table 3).

Table 14.1 Fraction of MTO households with different types of housing assistance

	Control group	Standard vouchers
Baseline		
Public housing	1.000	1.000
Interim evaluation (4–7 years after enrollment)		
Public housing	0.435	0.211
Housing vouchers	0.124	0.484
Privately owned subsidized project	0.099	0.062
No recorded assistance*	0.342	0.243
Final evaluation (10–15 years after enrollment)		
Public housing	0.296	0.186
Housing vouchers	0.252	0.446
Privately owned subsidized project	0.072	0.033
No recorded assistance*	0.380	0.335

*Some households with no recorded assistance almost surely received assistance from housing programs that are not required to report recipient addresses to HUD.

results of the final evaluation are also important for outcomes that depend on the cumulative effect of the treatment. They are far less important for other outcomes. For these outcomes, the entire time path of the results since the initial offer of housing vouchers is relevant. Due to attrition and crossovers, the differences between many outcomes for the two groups have gradually converged over time. However, if one alternative produces real benefits such as better housing in the early years of the experiment, that is highly relevant for judging its success even if this benefit eventually declines to zero over time. Better housing is valued for its own sake.

The simple difference in mean outcomes for the three groups is directly informative about the effect of an immediate voucher offer to residents of housing projects of the type studied compared with continuation of the current system. MTO research has been primarily concerned with the general issue of the effects of living in a better neighborhood, and most studies have focused on comparisons between the control group and the low-poverty voucher group. However, the two comparisons that are most directly relevant for housing policy are between the control group and the regular Section 8 voucher group and between the two treatment groups. If a reform along these lines is pursued, it will almost surely be to offer regular vouchers to public housing tenants. The second comparison is relevant for a reform of the current voucher program, namely, adding an additional restriction on voucher use. The MTO reports do not contain the information necessary to test hypotheses about the second reform. Therefore, this section will focus on the comparison between the control group and the regular Section 8 voucher group. We focus on ITT estimates because they are more relevant for policy analysis. They are estimates of the average effect of the intervention across the population that it is intended to help.

Orr et al. (2003) reported the results of the interim evaluation. Four to seven years after enrollment, the offer of a regular voucher led the treatment group to live in substantially better and safer neighborhoods and occupy somewhat better housing. The effects on all measures of adult physical health were small and statistically insignificant at the 5% level. The same was true for children with one exception. Boys between the ages of 12 and 19 in the treatment group were 50% more likely to have an injury requiring medical attention.¹⁰ The effects on all measures of adult mental health were small and statistically insignificant at the 5% level. Mental health outcomes were substantially better for girls and statistically significant at the 5% level in most cases. Mental health outcomes for boys were mixed and never statistically significant. Girls had better and boys had worse delinquency outcomes, but these effects were usually small and rarely statistically significant. The same was true for risky behaviors. The schools attended by the children were slightly better on a number of dimensions. However, the effects on educational outcomes were trivial. The same was true for adult earnings, employment, and participation in other welfare programs.

Although the mix of housing assistance received by the control and treatment groups differed greatly at the times of the interim and final data collection, MTO's final report paints a picture similar to the interim report (Sanbonmatsu et al., 2011). The regular voucher offer led members of the treatment group to live in better and safer neighborhoods and somewhat better housing. Girls and their mothers experienced better mental health and less obesity on some measures. These outcomes almost surely resulted from a reduction in stress from living in a safer neighborhood.

The official reports of the interim and final evaluations contain results for an enormous number of outcomes. Even if the true effects were zero, random sampling would almost surely generate statistically significant results for some outcomes. To avoid this problem, Kling et al. (2007) analyzed the results for five aggregates of 15 outcomes studied in the interim evaluation, namely, economic self-sufficiency, physical health, mental health, risky behavior, and education. Outcomes in the first category were for adults and those in the last two for youth. For adults, all of the estimated effects of the regular voucher offer were positive, but none were statistically significant at the 5% level. For female youth, all of the results were positive and the results for mental health and risky behavior were large and statistically significant. For male youth, most results were negative, and the negative effects on risky behavior and an aggregate of all outcomes were large and statistically significant. The specific reasons for gender differences in treatment effects are not well understood but seem to reflect differences in how male and female youths from disadvantaged backgrounds adapt and respond to similar new neighborhood environments (Kling et al., 2007, pp. 105–107).

¹⁰ Due to the large number of outcomes and subgroups studied, some outcomes of this type will result from random sampling even when the true effect is zero.

14.3.3.3 Experimental housing allowance program

No discussion of housing policy research would be complete without some mention of the EHAP. The major goals of EHAP were to determine the market effects of an entitlement program of household-based housing assistance and the effects of different types of such assistance on individual consumption choices. Congress authorized this program in 1970, planning for the experiment occurred in the early 1970s, data were collected during the mid-1970s, and the final reports were completed in the late 1970s and early 1980s. The experiment costs almost \$200 million (more than \$700 million in 2014 prices). Research and data collection accounted for almost half of this amount. The research firms that ran the experiments issued more than 300 reports, technical notes, and professional papers, and many others contributed to EHAP analysis.¹¹

The two largest and most important components of EHAP were the supply experiment and the demand experiment. The primary purpose of the supply experiment was to determine the market effects of an entitlement program of household-based assistance such as its effects on the market rents of units with specified characteristics and how suppliers alter their units in response to the program.¹² The experiment involved operating entitlement housing allowance programs in the Green Bay and the South Bend metropolitan areas. Eligible families (roughly the poorest 20%) were offered a cash grant on the condition that they occupy housing meeting certain standards. The poorest families were offered the largest grants. The demand experiment conducted in the Pittsburgh and Phoenix metropolitan areas was primarily intended to see how recipients would respond to different types of tenant-based housing assistance and, for a given type, to different program parameters. The most influential demand experiment research went beyond a comparison of different types of tenant-based assistance. It compared several outcomes of the minimum-standards housing allowance program with outcomes of the major established housing programs in existence at the time.

EHAP produced many results that have been influential in housing policy debates, though memory of these results has faded over time. One of the most important results of the supply experiment was that the entitlement housing voucher program tested had a minimal effect on the market rents and prices of housing units with unchanging

¹¹ The best introductions to this vast literature are the final reports of the supply experiment, the demand experiment, and the integrated analysis (Kennedy, 1980; Struyk and Bendick, 1981; Lowry, 1983); an edited volume containing summaries of the findings by the major contributors to EHAP research (Friedman and Weinberg, 1983); an edited volume containing evaluations of this research by outside scholars (Bradbury and Downs, 1981); a monograph containing some of the more technical results on consumer behavior from the demand experiment (Friedman and Weinberg, 1982); and U.S. Department of Housing and Urban Development's (1980) summary report.

¹² We refer to the assistance as household-based rather than tenant-based because homeowners were offered the same subsidy on the same terms as otherwise identical renters. About 40% of participants were homeowners.

characteristics. This allayed concerns that expanding the smaller nonentitlement Section 8 housing voucher program established in the early years of the experiment would have significant effects on the rents of unsubsidized units.

Another important finding of the supply experiment was that the program induced a substantial increase in the supply of units meeting the program's minimum standards. Despite the modest subsidies provided (about half of the taxpayer cost per recipient household of the Section 8 housing choice voucher program after accounting for inflation), the entitlement housing allowance program led to a 9% increase in the supply of apartments meeting minimum housing standards over its first 5 years. This resulted from upgrading the existing stock of housing (rather than the production of new rental housing) entirely in response to tenant-based assistance that required families to live in apartments meeting the program's standards in order to receive the subsidy.

Another influential finding of the demand experiment was that tenant-based assistance is more cost-effective than the types of project-based assistance that existed at the time. This result played an important role in persuading Congress to rely more heavily on tenant-based housing assistance to deliver housing subsidies.

14.3.4 Evidence on the performance of low-income housing programs

This section deals primarily with the evidence on the effect of each program or type of program compared with a counterfactual of no housing assistance, but it also covers some papers that explicitly compare the performance of different programs. [Olsen \(2003\)](#) surveyed the evidence on program performance through about 2000. The current chapter briefly summarizes this evidence and covers the more recent research in more detail.

14.3.4.1 Cost-effectiveness in providing housing

The largest and most consequential difference between housing programs is in their cost of providing equally good housing in equally desirable neighborhoods. The best evidence indicates that project-based assistance has a much greater cost than tenant-based assistance when it provides equally good housing broadly conceived.¹³ These studies define equally good housing to be housing that would rent for the same amount in the unsubsidized market in the same locality. This measure accounts for the desirability of the neighborhood as well as the housing itself. In the best studies, the estimated magnitude of the excess cost is enormous.

The best study of HUD's largest program that subsidized the construction of privately owned projects (Section 8 New Construction and Substantial Rehabilitation) found an

¹³ [Olsen \(2008, pp. 9–15\)](#) provided a detailed summary of the evidence on the cost-effectiveness of low-income housing programs. [Olsen \(2009\)](#) provided a detailed description and critical appraisal of the data and methods used.

excess total cost of at least 44% (Wallace et al., 1981). That is, the total cost of providing housing under this program was at least 44% greater than the total cost of providing equally good housing under the housing voucher program. This translates into excessive taxpayer cost for the same outcome of at least 72%. It implies that it would have been possible using housing vouchers to serve all of the people served by this program equally well and serve at least 72% more people with the same characteristics without any increase in public spending. The best study indicates even larger excess costs for public housing (Mayo et al., 1980). More recent evidence has confirmed the large excess cost of the Section 8 New Construction and Substantial Rehabilitation Program (Finkel et al., 1999, exhibit 5–1; Shroder and Reiger, 2000), and recent US Government Accountability Office (GAO, 2001, 2002) studies have produced similar results for the major active construction programs—LIHTC, HOPE VI, Section 202, Section 515, and Section 811. Burge (2011) found that tax credit developers capture most of the subsidy that the government provides on behalf of their tenants, indicating some combination of excess profits, rent-seeking expenditures, and distortions in input usage. Eriksen (2009) provided some evidence about the latter.

The preceding evidence on the cost-effectiveness of project-based assistance applies to units built or substantially rehabilitated under a subsidized construction program and still under their initial use agreement. Evidence from the Mark-to-Market program indicates the excessive cost of renewing use agreements for privately owned subsidized projects (Hilton et al., 2004), and the EHAP provides clear evidence on the superior cost-effectiveness of tenant-based versus project-based housing vouchers (Mayo et al., 1980, pp. 134–139).

In contrast, a succession of studies over the years have found that the total cost of various types of tenant-based housing assistance has exceeded the market rent of the units involved by no more than the cost of administering the program (Mayo et al., 1980; Wallace et al., 1981; Leger and Kennedy, 1990; ORC/Macro, 2001, chapter V). That is, landlords receive market rents for their units.

There is no good evidence on the importance of various reasons for the large differences in the total cost of providing equally good housing under programs of tenant-based and project-based assistance. Among the plausible explanations are the absence of a financial incentive for good decisions and monitoring employee performance on the part of civil servants who operate public housing, the excessive profits that inevitably result from allocating subsidies to selected developers of private subsidized projects, the resources that developers devote to securing the limited subsidies available, and the distortions in usage of inputs resulting from the subsidy formulas. A special case of the latter is that project-based assistance is usually tied to the construction of new units. The least-expensive approach to improving the housing conditions of low-income households involves heavy reliance on upgrading the existing housing stock, the primary mechanism through which tenant-based assistance achieves this goal.

The results concerning the cost-effectiveness of different housing programs illustrate the virtue of forcing sellers to compete for the business of buyers. Under a program of tenant-based assistance, only suppliers who provide housing at the lowest cost given its features can remain in the program. If the property owner attempts to charge a voucher recipient a rent in excess of the market rent, the tenant will not remain in the unit indefinitely because he or she can move to a better unit without paying more for it. Under programs of project-based assistance, suppliers who receive payments in excess of market rents for their housing can remain in the program indefinitely because their tenants would lose their subsidies if they moved. These suppliers have a captive audience.

Although the weight of the evidence is substantial, all of the studies have some methodological problems, and none make highly accurate estimates of all of the magnitudes required to implement their methodology. The evidence is most deficient for some of the largest and fastest growing programs. In light of the results of existing studies and the consequences of using highly inefficient programs to deliver housing subsidies, cost-effectiveness studies of all of the major discretionary expenditures on project-based housing assistance such as incremental commitments under the LIHTC and HOME programs, renewals of use agreements with owners of privately owned subsidized projects, public housing operating and modernization subsidies, and project-based Section 8 vouchers should be the highest priority for housing policy research.

14.3.4.2 Housing consumption

The primary purpose of low-income housing assistance is to induce recipients to occupy better housing than they would choose if given an equally costly unrestricted cash grant. The literature is deficient in showing whether this goal has been achieved. Few studies compare the effects of housing programs with the effects of cash grants, and no recent study estimates the effect of any housing program on a comprehensive measure of housing consumption.

Many older studies estimated a program's effect on a comprehensive measure of housing consumption. When dealing with data for a single housing market, they used the market rent of the subsidized unit as an index of its overall desirability. When dealing with data from many housing markets, market rent was divided by an index of the rent of identical units across markets. This measure accounts for differences in the size, amenities, condition, and neighborhood of the units as well as their convenience to jobs, shopping, and recreation.

These studies require estimates of the market rent of each subsidized unit and the market rent of the unit that its recipient would have occupied in the absence of housing assistance. The estimated market rent of a subsidized unit was almost always based on a hedonic regression model estimated with data on unsubsidized rental units. The hedonics differed greatly with respect to the list of housing and neighborhood characteristics included. Most

were based on detailed information about the characteristics of the housing itself similar to the information in the American Housing Survey. However, the modest information about neighborhood amenities and convenience to jobs, shopping, and recreation in these hedonics might well have led to significantly biased predictions of the market rents of subsidized units, especially in public housing projects that tend to be located in the worst neighborhoods. In the older studies, estimates of the market rent of the unit that a recipient would have occupied in the absence of housing assistance were almost always based on the average housing expenditure of unsubsidized households with the same observed characteristics. Self-selection and administrative selection in program participation were ignored in most studies.

Olsen (2003, table 6.8) summarized the results of 8 studies that produced 18 estimates for 4 different housing programs. Almost all are based on data from the 1970s. The nine estimated percentage increases in mean housing consumption for public housing tenants ranged from 22% to 82% for different times and places, the four estimates for occupants of HUD's privately owned subsidized projects ranged from 26% to 58%, and four for voucher recipients ranged from 16% to 63%. These results could be far from the mark today, especially for subsidized projects because they are much older. However, Walters (2009) presented evidence suggesting that later subsidies for modernizing these projects have offset their deterioration to a considerable extent.

More directly relevant for assessing the performance of low-income housing programs are the results of four studies that estimate how much (if at all) they increased housing consumption beyond what would result from unrestricted cash grants in amounts equal to each recipient's housing subsidy (that is, the difference between the market rent of its unit and the tenant's rent). The four estimates for public housing based on data from 1965 to 1984 indicate that this program increased mean recipient housing consumption from 40% to 53% beyond what would have resulted from cash grants in these amounts. The one estimate for the voucher program based on data for 1976 indicated only a 10% increase beyond cash grants. The only estimate for the entire system of low-income housing assistance showed a 39% increase beyond cash grants based on 1977 data. These results indicate that at these times, housing programs changed consumption patterns in ways consistent with their traditional rationales.

Updated results for current programs are long overdue. There have been few studies of the effects of low-income housing programs on housing consumption in the past three decades, and these studies are limited to a few housing characteristics or recipient ratings of their housing. They do not attempt to determine whether the programs induce recipients to occupy better housing than they would choose if they were given unrestricted cash grants in the amounts of their housing subsidies. Furthermore, an alternative calculation for programs of project-based housing assistance would be of considerable interest. The cost incurred by taxpayers on behalf of occupants of subsidized housing projects exceeds the tenant subsidy by much more than the administrative cost of the programs involved due to their

cost-ineffectiveness. If these tenants had been given cash grants in amounts equal to the taxpayer cost minus the program's administrative cost, their housing consumption with cash grants would have been much larger, and it is entirely possible recipients would have consumed more housing services and other goods than under the housing program. The effect of this alternative to project-based housing assistance has not been estimated.

Using data from the 1990–1995 Current Population Survey (CPS) and the 1990 decennial census and two-stage IV estimation to account for the endogeneity of program participation, [Currie and Yelowitz \(2000\)](#) found that living in a subsidized housing project reduces overcrowding, one aspect of the housing bundle. This is not surprising because programs that subsidize housing projects have rules to ensure that families have adequate space. Larger families are assigned to units with more bedrooms. Because CPS respondents said that they lived in a public housing project, the authors assume that their results apply to the public housing program. However, data from the 2011 American Housing Survey that identifies the type of HUD assistance received by each household from administrative records as well as the respondent's answer to the CPS question reveal that about 56% of the people who say that they live in public housing projects live in privately owned subsidized projects. Almost all of the rest live in public housing. Therefore, Currie and Yelowitz's results apply to subsidized housing projects as a group.

The random assignment voucher evaluation provides the only recent evidence on the effects of HUD's rental housing programs on housing consumption. As explained in the section that describes the experiment in detail, its TOT estimates are reasonably viewed as evidence of the effect of voucher usage compared with no assistance. The study found that voucher usage decreased its measure of crowding by 44%, increased the fraction of recipients who rated the condition of their housing as good or excellent by 8 percentage points, and decreased the fraction of recipients who reported at least two of seven specific problems with their units by 7 percentage points ([Abt Associates Inc. et al., 2006](#), exhibit 5.3). The latter two impacts were not statistically significant at the usual levels. No recent evidence is available for HUD's programs of project-based assistance.

[Burton et al. \(2000, table 8\)](#) provided the only evidence on the effect of the tax credit program on housing consumption. It is based on the opinions of 839 occupants of 39 tax credit projects. Overall, 54% considered their tax credit apartment better than their previous unit, 24% considered it about the same, and 22% considered it to be worse. The respondents who considered the apartment itself to be worse presumably benefited in other respects such as lower rent, better neighborhood, or more convenient location. Unfortunately, the study does not provide a quantitative index of the extent of the increases and decreases in the overall housing consumption of these families. So it is not known how the average magnitude of the improvement among families who moved to better housing compared with the average magnitude of the reduction in the desirability of the housing among those who moved to worse housing. In assessing this evidence on the effect of the tax credit program on housing consumption, it is important to

realize that none of the units involved were more than 7 years old. Over time, the condition of these units will surely deteriorate. What is relevant for assessing the effect of the tax credit program on housing consumption is a comparison of the housing provided by the tax credit units over the entire period that their use is restricted to low-income families with the housing that their occupants would have occupied in the absence of this program. Since the use agreements extend for at least 30 years, it seems likely that these results from the early years of the projects overstate the average effect of the program on housing consumption over the period that their rents are limited by the agreement.

One aspect of housing consumption of particular policy interest is homelessness. If anyone should receive housing assistance as opposed to an unrestricted cash grant, it is homeless people. Some have serious mental illnesses that impair their judgment. Others impose tangible external costs on others.

Abt Associates Inc. et al. (2006, exhibit 5.3) estimated that receipt of housing vouchers reduced the number of families that would have been on the streets or in a shelter at some point during the year prior to final data collection by about 9 percentage points. Since receipt of housing assistance precludes homelessness, this is also their estimate of the number of recipients that would have been on the streets or in a shelter at some point during the prior year. Because participants were single women with children, most surely would have been in shelters. Data on this matter were collected but not presented. Other types of housing assistance would surely have a similar effect on the same types of households. Presumably, the programs that serve the poorest have the largest effect. However, this has not been studied for other programs.

Abt's study deals with the effect of a particular housing program on homelessness sometime during the past year for a subset of families eligible for housing assistance. Other studies estimate the effects of the entire system of low-income housing assistance on homeless at a point in time for all recipients. Several studies estimate the fraction of households receiving low-income housing assistance that would have been homeless in its absence and the effect on homelessness of reforms of existing programs and implementation of new programs. Using separate data sets on housed and homeless individuals and choice-based sampling methods, Early (1998, 2004) found that at most, 3.8–5.0% of assisted households would have been homeless in the absence of housing assistance.¹⁴ Early (2004, table 1) also reported that the mean *per capita* household income of the homeless is less than a half of the mean for households that receive housing assistance. Using data from a systematic count of the homeless throughout the United States, Early and Olsen (2002) found that a sufficiently large increase in the targeting of existing housing assistance on the poorest people would eliminate homelessness among those who

¹⁴ It is important to realize that most of the counted homeless have roofs over their heads each night. They live in a special type of subsidized housing called homeless shelters, and the best shelters provide good housing. It is also important to realize that the street homeless are undercounted.

apply for assistance. The main residual problem would be persuading mentally ill people living on the streets to accept the help available.

In a paper that calibrates the Anas and Arnott general equilibrium model of the housing market with data from Los Angeles, San Diego, San Francisco, and Sacramento, [Mansur et al. \(2002\)](#) found that an entitlement program that provides a cash subsidy to all renters in the lowest income quintile would reduce homelessness by between 25% and 33% across the four areas.¹⁵ This is obviously a less targeted approach to preventing homelessness than housing assistance to the very poorest households, and it would cost five times as much as current spending on housing assistance ([Early, 2004](#), p. 199).

14.3.4.3 Neighborhood occupied

Neighborhood amenities are important parts of the housing bundle broadly conceived, and many recipients of housing assistance move to different neighborhoods. Older studies of this effect usually compared the neighborhoods of participants immediately before and after entering the program ([Olsen, 2003](#), pp. 407–411). Generally speaking, this evidence indicates that most low-income housing programs have a modest effect on the types of neighborhoods occupied. Public housing is the exception. Its neighborhoods were noticeably poorer than those previously occupied.

More recent studies use different methods but reach similar conclusions. For example, [Susin \(2005, p. 207\)](#) found that public housing tenants live in census tracts with poverty rates 8.8 percentage points higher than in the absence of assistance, tenants in HUD-subsidized privately owned projects live in tracts with poverty rates 2.6 percentage points higher, and voucher recipients live in tracts with poverty rates 2.3 percentage points lower. Using excellent data and methods described later, [Carlson et al. \(2012a\)](#) and [Jacob and Ludwig \(2012\)](#) found that housing vouchers lead their recipients to occupy neighborhoods that are only slightly better in a variety of dimensions. [Eriksen and Ross \(2013\)](#) produced similar results based in part on data from the voucher evaluation and IV estimation using assignment to its voucher treatment group as an instrument for voucher usage. Their results suggest that many voucher recipients initially use their voucher in their current unit or immediate neighborhood in order to secure the voucher and then move to a better neighborhood later.¹⁶ The secondary moves are typically to better neighborhoods, though not markedly better. The final report of the voucher

¹⁵ Because the model divides all housing units into quartiles based on market rent and assumes that units within a quartile are identical in their rent and desirability, a household must occupy a unit better than the worst in order to receive a subsidy. The mean subsidy ranged from \$2708 per household (in 2014 prices) in Sacramento to \$4914 in San Francisco. The poorest received much greater subsidies because the subsidy was equal to a payment standard –30% of household income. Unlike experience with actual entitlement programs, all eligible households were assumed to receive assistance.

¹⁶ Voucher recipients have between 60 and 120 days to move into a unit meeting the program's minimum housing standards, and many already live in units meeting these standards.

evaluation paints a similar picture based on different self-reported measures of neighborhood desirability, namely, five measures of criminal victimization, one measure of visible drug use or sales, and five other measures ([Abt Associates Inc. et al., 2006](#), exhibit 3.8).

[Buron et al. \(2000, exhibit 3–17\)](#) provided evidence for the tax credit program based on recipient opinions. About 19% of respondents who moved into a tax credit project were from the same neighborhood. Almost equal numbers of respondents who came from another neighborhood considered the project's neighborhood to be better, about the same, and worse overall than their previous neighborhood. Respondents were also asked about 10 specific aspects of their neighborhoods. In 8 of the 10 aspects, a plurality said that their previous and new neighborhoods were about the same. In 9 of 10, more said that the neighborhood was better than said it was worse. The percentage of households who considered their tax credit neighborhood better than their previous neighborhood varied from 40% for access to public transportation to 26% for access to good schools. Strikingly, slightly more respondents thought that the neighborhood of their tax credit project was a worse rather than a better place for children than their previous neighborhood.

[Lens et al. \(2011\)](#) shed new light on the differences in the neighborhoods occupied by voucher recipients, public housing tenants, and occupants of tax credit projects. Like [Newman and Schnare \(1997\)](#), they found that public housing tenants live in census tracts with markedly higher poverty rates than participants in the other two programs and that the difference between voucher recipients and occupants of tax credit projects in this regard is modest. The picture is very different for neighborhood crime rates. Occupants of tax credit projects live in neighborhoods with crime rates about 30% greater than voucher recipients and only slightly less than the crime rates of the neighborhoods of public housing tenants. Since voucher recipients have much more choice concerning the location of their housing, this suggests that housing projects are poorly located from the viewpoint of recipient preferences. In MTO, getting away from crime was by far the most important reason offered by participants for wanting to move from their public housing project ([Orr et al., 2003](#), exhibit C1.3). In about 43% of enrolled households, some member had been a crime victim in the previous 6 months ([Sanbonmatsu et al., 2011](#), exhibit 1.2).

To understand the pattern of the results, it is important to recognize the ability of local political bodies to influence the location of housing projects. For public housing, they had a veto over location. Due to neighborhood opposition, public housing projects, especially for families with children, were usually built in the poorest neighborhoods. For privately owned subsidized projects, local political bodies can exert considerable influence through land use regulations, for example, through denying or granting variances from zoning ordinances. Furthermore, these programs often provide incentives for developers to choose particular locations. Their proposed location can affect the probability that they will be selected to receive a subsidy ([Gustafson and Walker, 2002](#)). It can

also affect the magnitude of the subsidy received. For example, the tax credit program offers subsidies that are almost 30% greater for projects located in qualified census tracts. These are areas with unusually large numbers of the poorest families. Using RD methods, [Baum-Snow and Marion \(2009\)](#) showed that this feature has a large effect on the location of tax credit projects. Other program features have important, but less obvious, effects on project locations. For example, the rent ceiling in tax credit projects is the same at all locations in a metropolitan area and the tax credit subsidy does not depend on land cost. This gives developers an incentive to avoid neighborhoods with high land prices ([Lang, 2012](#)). In short, local political bodies and private developers decide where subsidized housing projects are built. Due to their below-market rents, recipient preferences have less impact than in the private market.

14.3.4.4 Racial and economic integration

The moves that result from housing assistance lead to a different degree of economic and racial integration in the broader housing market. Many people care deeply about these outcomes.

The usual approach to determining whether a program contributes to racial or economic integration has been to compare the characteristics of the preprogram neighbors of recipients with their neighbors under the program. For example, if 20% of the people in a metropolitan area are black, the mean percentage black in the preprogram neighborhood of black recipients is 60%, and the mean percentage in the neighborhood occupied under the program is 30%, the conclusion of this approach is that the program has reduced racial segregation.

This approach fails to account for the chain of moves that result from the initial moves of recipients. A simple example illustrates the problem. Suppose that an existing apartment building is demolished to build a new subsidized project with the same number of dwelling units. Assume that this neighborhood had the metropolitan mean percentage black households, say, 20%. Suppose that the new building has the same fraction black households as the old building. Assume that the blacks moving into the neighborhood came from neighborhoods with more than 20% black and the whites came from neighborhoods more than 80% white. The usual approach would conclude that the housing program promoted racial integration. However, this is not necessarily the case. For example, if the whites displaced by the demolition of their building moved into the units vacated by whites moving into the project and the displaced blacks moved into the units vacated by blacks moving into it, there would be no effect on the degree of racial segregation.

Recent studies have begun to address the deficiencies of the standard approach. [Baum-Snow and Marion \(2009\)](#) showed that tax credit projects induce some initial homeowners in their neighborhoods to move and attract households with lower

incomes. [Horn and O'Regan \(2011\)](#) explored various mechanisms through which tax credit projects might affect racial segregation and found that the tax credit program has reduced racial segregation at the metropolitan level.

Recent advances in the specification and estimation of equilibrium models of housing demand and neighborhood choice ([Epple and Sieg, 1999](#); [Bajari and Kahn, 2005](#); [Bayer et al., 2007, 2011](#); [Bayer and McMillan, 2012](#)) together with the increasing availability of data on the addresses of recipients of housing assistance will make it possible to learn more about the effects of low-income housing programs on economic and racial segregation.

In one of the first applications, [Galiani et al. \(2012\)](#) used MTO data to estimate such a model and used the estimated model to analyze the effect of alternative restrictions on where voucher recipients must use vouchers in the first year. They found that limiting the initial use of the voucher to a census tract with a poverty rate less than 5% rather than 10% would reduce voucher take-up so much that it would increase the poverty rate of the neighborhoods occupied by families offered vouchers. A larger number of families would remain in their initial high-poverty neighborhoods. Loosening the restriction to neighborhoods with poverty rates less than 20% rather than 10% would *reduce* the mean poverty rate among households offered vouchers slightly but increase voucher usage greatly (from 39% to 58%) and increase the value of the program to households offered vouchers. They also found that adding specific constraints concerning the racial composition of the neighborhoods that could be occupied during the first year would reduce take-up but have little effect on economic or racial segregation.

14.3.4.5 Consumption of other goods

Neither recipients nor taxpayers care only about recipient housing consumption. The creation of large programs to subsidize the consumption of other goods for low-income households indicates taxpayer interest in recipient consumption of these goods, and most US housing programs have been designed to increase overall consumption of other goods (usually described as increasing the affordability of housing). Many early studies of the effects of low-income housing programs estimated their effect on overall consumption of nonhousing goods and services, that is, expenditure on these goods divided a price index. [Olsen \(2003, table 6.8\)](#) summarized the results of four studies that produced eight estimates for three programs prior to 2000. Almost all are based on data from the 1970s. The six estimated percentage increases for public housing ranged from 5% to 19% for different times and places. The only study of HUD's privately owned subsidized projects indicated that it had no effect on consumption of other goods. However, at the time, this program had an unusual feature, namely, a substantial minimum rent, that surely makes this result unrepresentative of HUD's programs of this type. These programs almost surely have increased overall recipient consumption of nonhousing goods and services in most time periods. The only study of the housing voucher program indicated that it increased consumption of other goods and services by 50%. As in the case of older

studies of the effects on overall housing consumption, almost all of these studies ignore self-selection and administrative selection in program participation.

Almost all studies of the effects of low-income housing programs on consumption patterns have assumed that they had no effect on earnings. That is, they have assumed that observed earnings under the program are the same as what earnings would have been in its absence. Because these programs reduced earnings, these studies have understated recipient consumption of housing services and other goods in the absence of the program and hence overstated the program's effect on consumption of both composite commodities.

Recent evidence on the effect of housing programs on nonhousing consumption is meager. The voucher evaluation found that receipt of vouchers reduced housing expenditure by 40% (Abt Associates Inc. et al., 2006, exhibit 5.3). The resulting percentage increase in consumption of other goods is not reported. However, based on the information provided, it was roughly 50%. Given the similarity of the economic circumstances of their participants and their formulas for determining tenant rent, public housing and HUD's programs that subsidize privately owned projects are likely to have had roughly similar effects. However, there are no recent estimates.

Unlike almost all other low-income housing programs, the tax credit program appears to have had little effect on the housing expenditure of tenants and hence their consumption of other goods. Buron et al. (2000, exhibit 3–9) indicated that 47% of the occupants of tax credit projects paid lower rents for their tax credit apartments than their previous apartments, 13% paid about the same amount, and 40% paid more. Since few of these households had just moved into their tax credit units and rents of identical units increased continuously over the period, these results almost surely overstate the extent to which these households paid lower rents than in the absence of the tax credit program.

14.3.4.6 Mean tenant benefit and subsidy

Since many economists are interested in the extent to which housing subsidies differ from lump-sum grants, many studies compare the recipient benefit with the subsidy. The most common measure of recipient benefit used in studies of housing programs is the equivalent variation, that is, the lump-sum grant that is just as satisfactory to the recipient as the housing program.¹⁷ As mentioned earlier, the tenant's subsidy is the excess of the market rent of the unit occupied over the rent paid by the tenant.¹⁸ Since most housing programs change budget spaces in ways very different from lump-sum grants, we certainly expect any satisfactory measure of benefit to be less than the subsidy for almost all participants. In

¹⁷ Hammond (1987, chapter 2) provided a detailed account of the evolution of benefit estimation for low-income housing programs.

¹⁸ The subsidy is less than the taxpayer cost for all programs due to their administrative costs. It is much smaller for programs of project-based assistance due to their cost-ineffectiveness.

other words, we expect almost every household to consume a different bundle of goods than it would choose if it were given a cash grant equal to its subsidy. Therefore, the mean benefit should be less than the mean subsidy.

Olsen (2003, table 6.17) summarized the results of the seven studies that estimate both the mean benefit and subsidy for one program or for the entire system of low-income housing assistance. Estimated benefits are based on estimated indifference maps or equivalently estimated demand functions. There are few estimates for programs other than public housing, and there are no recent estimates for any program or for the system as a whole. The median of the estimated ratios of mean benefit to mean subsidy for public housing is 0.76, and 70% of the estimated ratios are between 0.71 and 0.81. Based on one study apiece, the ratio is between 0.63 and 0.77 for HUD's largest program that subsidized privately owned projects (Section 8 New Construction and Substantial Rehabilitation), 0.83 for the housing voucher program, and 0.61 for the system as a whole.

All of these studies almost surely overstate tenant benefits. They are based on the implicit assumption that each recipient of housing assistance occupies his or her most preferred dwelling unit among all units with the same market rent as the unit occupied under the program. Due to the minimum housing standards, this isn't true even for voucher recipients. It is surely much further from the truth for families living in subsidized projects. They were given an all-or-nothing offer of a particular dwelling in a particular location. No published study of US low-income housing programs has addressed this important distortion. Glaeser and Luttmer (2003) analyzed a similar distortion in several elements of the housing bundle for rent control in New York City. Lui and Suen (2011) did it for one element for public housing in Hong Kong.

In assessing the significance of these results for public policy, it is important to realize that mean recipient benefit will be less than the mean subsidy for any successful housing program. To account for the views of nonrecipients who pay for them, housing programs are intended to achieve outcomes different from unrestricted cash grants.

14.3.4.7 Distribution of benefits

Many nonrecipients care about how benefits are distributed across recipients of housing assistance—both how mean benefit varies with household characteristics and the variance in benefit among households with the same characteristics. Most of the older studies that estimated tenant benefits also explored these matters.

The design of housing programs does not ensure that mean benefit will vary with household income within a locality in a particular manner. Most recipients of low-income housing assistance in the United States (except those in tax credit projects) pay 30% of their adjusted income in rent. Therefore, if all households of a particular size served by a program lived in apartments with the same market rent, the subsidy would be larger for the poorest households of that size. Under all variants of tenant-based housing vouchers, the maximum subsidy has varied inversely with income among households of

the same size. However, it is far from the truth that all households served by a mature construction program (that is, a construction program that has been in existence for many years) occupy housing with the same real market rent. Furthermore, a larger subsidy does not imply a larger benefit. If poorer households experience a larger consumption distortion than richer households, they could receive a smaller benefit even though they receive a larger subsidy.

Olsen (2003, table 6.18) summarized the results of regressions of estimated benefit on household characteristics in which a linear relationship between mean benefit and income, family size, age, race, and other characteristics is assumed. Some results are consistent across the studies. For public housing, housing vouchers, and the system as a whole, the mean benefit is larger for poorer households that are the same with respect to other characteristics. Similarly, mean benefit is larger for larger households. These results continue to hold when the authors allow for the possibility of a nonmonotonic relationship by including income and family size squared. The results are less consistent for race and age of the head of the household. The coefficients have different signs in different studies, they are often statistically insignificant, and the magnitudes of the coefficients indicate small differences in mean benefit among otherwise similar households who differ in these respects.

Several other noteworthy results emerged from these analyses. First, several studies have found substantial differences across geographic areas in the real mean benefit of public housing, housing vouchers, and the entire system of housing subsidies among families with the same real income and other demographic characteristics (Murray, 1975; Reeder, 1985; Hammond, 1987). Second, when Reeder included in the regression equation a measure of the household's taste for housing based on information on its preprogram consumption pattern and budget constraint, he found that households with the strongest taste for housing received the largest benefit from the voucher program. Unlike simple subsidies that reduce the price of a good and allow recipients to choose its quantity, this result is not inevitable. It results from the program's parameters and the joint distribution of recipient tastes and incomes.

14.3.4.8 Labor earnings and employment

The perennial desire to help the poorest people has always been combined with a desire to avoid their prolonged dependence on others. For people expected to work, the ideal has been to help them help themselves by increasing their productivity. Increasing the future productivity of children in low-income households has been a particular interest. President Lyndon Johnson promoted the War on Poverty with the phrase "a hand up, not a hand out." President Clinton proposed and implemented major welfare reforms "to make work pay."

Despite the welfare reforms that occurred during the 1990s such as increasing the generosity of the earned income tax credit and replacing the Aid to Families with Dependent

Children (AFDC) program with the TANF program that imposed time limits and other strong incentives to promote market labor supply, many programs that provide assistance to low-income households reduce the amount of assistance as labor earnings increase, and many low-income households have participated in multiple programs of this sort and faced extremely high reductions in their benefits with increases in their earnings.

Most low-income housing programs in the United States reduce their subsidy by 30 cents for each additional dollar of countable income. Economists have focused on this feature of the subsidy formula in thinking about the likely effects of these programs on the labor earnings of adults in assisted households. The standard model of consumer choice implies that a program of cash assistance of this type will induce its recipient to earn less. Standard economic theory does not, however, have an unambiguous implication about this effect for housing programs with the same subsidy formula. When account is taken of the restrictions on housing consumption in low-income housing programs, it is consistent with general theory that an individual would increase labor supply in response to these programs (Schone, 1992).

Furthermore, some have suggested other mechanisms through which housing assistance might increase labor earnings of adult recipients over the long run (Abt Associates Inc. et al., 2006, pp. 82–84). For example, housing assistance may induce recipients to live in safer neighborhoods where they would enjoy better mental and physical health. Similar arguments apply with greater force to the future earnings of children in recipient households. Housing assistance might lead them to become more productive adults. The disincentives resulting from the subsidy schedule are only relevant to children in recipient households to the extent that they think ahead and expect to receive low-income housing assistance later in life. In this case, the subsidy schedule reduces their incentive to devote efforts to increasing their potential earnings.

The effects of cash assistance programs on labor earnings and employment have been heavily studied for quite some time. Research on the effects of in-kind transfers has been much slower to develop. However, in the late 1990s and early 2000s, research on these effects of low-income housing programs expanded rapidly. Shroder (2002b) cited 18 papers on the short-term effect of housing assistance on employment and earnings that were completed during this period. The results of these studies are mixed. The majority of studies found that housing assistance decreased earnings and employment. However, some indicated the opposite effect. Most estimated effects were relatively small, and hypothesis tests often failed to reject the hypothesis of no effect at standard levels of significance. Generally speaking, the data used in these studies were poorly suited to studying labor supply effects. Strong assumptions were required to justify the interpretation of the results as causal effects. As a result, they left considerable uncertainty about the magnitude and even the direction of the effect.

This situation has improved markedly in the past decade. Three studies of HUD's largest low-income housing program (the housing voucher program) have assembled

data well suited to studying its labor supply effects on adult recipients and used excellent statistical methods to analyze it. Other good studies have produced estimates of these effects for programs of project-based housing assistance as well as housing vouchers. Although the results of the studies don't agree in every detail, the big picture is clear. US low-income housing programs induce adult recipients to earn 10–15% less on average. The evidence indicates that the reduction in the magnitude of the subsidy that results from higher labor earnings is an important reason for this effect. For example, the evidence indicates that the three broad types of housing assistance have similar effects on labor earnings and employment. These programs differ in respects that some believed would significantly affect labor earnings. What they have in common is their subsidy formula.

Jacob and Ludwig (2012) provide the best evidence on the effects of housing vouchers on labor supply. It is based on the choices of 11,696 households offered housing vouchers in Chicago between 1997 and 2003 and 30,662 households who applied for vouchers but were told that they would not receive them. In 1997, the organization that administered the voucher program opened its waiting list for enough time to allow over 82,000 eligible households to get on it. This greatly exceeded the number of vouchers that would be available anytime soon. To ration the available vouchers, the organization randomly assigned each household a position on the list. About 18,000 households were offered vouchers by this organization between 1997 and 2003. Over 47,000 were so far down the list that they were told that they would not receive vouchers. The households in these two groups that were headed by working-age, able-bodied adults who did not receive housing assistance at baseline formed the control and treatment group for the analysis. Jacob and Ludwig assembled information about many baseline characteristics of these households, their quarterly labor earnings and receipt of public assistance (AFDC/TANF, food stamps, and Medicaid), and their addresses from 6 years prior to receipt of assistance through 8 years after receipt.

Their analysis of this impressive data set indicated that receipt of a housing voucher reduced labor force participation by 6% and labor earnings by 10% on average. These are TOT effects where treatment is defined as usage of the voucher offered in at least the first year after it was offered. A household was viewed as treated even after it relinquished its voucher.

Jacob and Ludwig also studied how the ITT effect varied over time since the initial offer. Their results indicated that the negative effects on earnings and employment increase over time. This result was surprising because some initial voucher recipients relinquished their vouchers each year and there are some reasons to expect that housing assistance would have positive long-run effects on labor earnings and employment. One possible explanation is that the reduction in hours worked and the acceptance of a less challenging job during the period of receipt of housing assistance reduces a person's future wage rate. Working itself increases skills through repetition and on-the-job training. Reduced labor force

engagement during the time that a person receives low-income housing assistance might also send a signal to future employers about the person's energy and enthusiasm for a steady job. These factors may more than offset the effects of attrition and other factors.

Finally, Jacob and Ludwig explored two mechanisms that have led others to believe that housing assistance would increase recipient earnings and employment, at least after an initial adjustment period, namely, that it would promote residential stability and induce recipients to move to neighborhoods closer to jobs appropriate for their skills and with neighbors who would be helpful in finding them. Jacob and Ludwig found that vouchers have very small effects on neighborhood choices and residential stability. This leads them to attribute their estimated work disincentive effects to the subsidy formula.

Two other excellent studies produced similar results concerning the immediate effect of vouchers on earnings and employment. The evaluation of the Welfare to Work Voucher program found that receipt of housing vouchers reduced labor supply about 12% in the first 6 months (Abt Associates Inc. et al., 2006, exhibit 4.9). Carlson et al. (2012b) produced the same result for the first year of voucher receipt based on another excellent data set. Their sample consisted of all households in Wisconsin in 2001, 2002, and 2003 that did not live in public housing but received TANF or food stamp benefits. Each year, some households were new voucher recipients and others did not receive vouchers. The authors match each voucher recipient with five nonrecipients with the closest propensity scores based on many characteristics, including earnings and employment in the previous 5 years. Difference-in-difference estimators are used to produce the results from this sample.

Like Jacob and Ludwig, Abt et al. and Carlson et al. rejected a number of mechanisms through which voucher receipt has been hypothesized to affect earnings and employment. Abt et al. found little effect of voucher receipt on travel time to work or changed transportation mode (p. 107), finding a job through a neighbor (p. 110), or increased education and training (p. 117). Carlson et al. (pp. 136–137) found that voucher receipt has little effect on the type of neighborhood occupied. This leads Carlson et al. (p. 143) to attribute its negative effect on labor earnings and employment to the subsidy formula. Abt et al. (p. 128) attributed it mainly to the temporary disruption caused by searching for housing.

Several differences between the three studies should be mentioned. First, because the policy interest in work disincentive effects pertains to working-age, nondisabled adults, Jacob and Ludwig and Abt et al. restricted their main analysis to this group, and the preceding results refer to it. Carlson et al.'s main results are based on a sample that includes elderly and disabled adults, and the housing voucher program serves many such people. Their estimated work disincentive effects would have been larger if their analysis had been limited to working-age, nondisabled adults. This is clear from their own subgroup effects for different ages and Jacob and Ludwig's supplementary analyses for elderly and disabled (table 4).

Second, the samples used by Abt et al. and Carlson et al. are likely to understate the average work disincentive effects of the housing voucher program across all recipients. About 80% of Abt et al.'s sample received TANF benefits at baseline (p. 24), and these benefits are included in the voucher program's countable income. The voucher subsidy is reduced by 30 cents for each additional dollar of countable income. For households without cash assistance, the program reduces the subsidy by 30 cents for each additional dollar of earnings. This is not the case for TANF recipients. TANF reduces its benefits with increases in the family's earnings. As a result, an increase in earnings does not lead to a reduction of 30 cents in the voucher subsidy for each additional dollar of earnings. In the extreme where the benefit reduction rate in TANF is 100%, increased earnings would have no effect on the housing program's countable income and hence its subsidy. If the benefit reduction rate in the cash assistance program were 50%, participation in the housing program would add 15 rather than 30% to the overall benefit reduction rate because a dollar increase in earnings would add only 50 cents to a housing program's countable income. Nominal benefit reduction rates in US cash assistance programs have typically ranged between 50% and 100%.¹⁹ In short, housing assistance adds less to the overall benefit reduction rate of TANF recipients than other voucher recipients. Not surprisingly, Jacob and Ludwig (2012, table 4) found that voucher receipt has a much smaller effect on labor force participation of households that received TANF benefits at baseline.

This is much less of a problem for Carlson et al.'s analysis because their sample is based on receipt of TANF or food stamp benefits, and less than a 10th of this group participated in TANF. Unlike cash assistance, food stamp benefits are not included in the countable income that determines the generosity of the housing voucher subsidy. For households that receive food stamps and other in-kind assistance but not cash assistance, housing assistance adds 30 percentage points to their benefit reduction rate.

Unlike Jacob and Ludwig, Carlson et al. and Abt et al. found that the work disincentive effects fade over time, albeit at different rates. Carlson et al.'s results suggest that the fade-out takes more than 5 years; the Abt et al. results suggest 3 years. One obvious reason for the estimated fade-out of the effect of low-income housing assistance in these studies is that some households with housing assistance give it up each year. The fraction of these households that continue to receive housing assistance declines each year. Since the housing program's benefit reduction rate is no longer relevant to the households that have left the program, a fade-out in the estimated effect should not be surprising. The effect of the program on a person eventually fades out because the person does not remain in the program forever. Only the presence of other forces working in the opposite direction would prevent it.

¹⁹ Abt Associates Inc. et al. (2006, p. 83) reported TANF benefit reduction rates in this range at their sites.

Since the treatment groups in all three studies experience departures from the voucher program, the puzzle is why Jacob and Ludwig's estimated reduction in earnings and other effects get larger with the passage of time from initial receipt of assistance. One possible reconciliation is that the reduction in hours worked and the acceptance of a less challenging job that results from receipt of housing assistance and reduces a person's future wage rate are more important for people who don't participate in TANF. Whatever the explanation, Jacob and Ludwig argued that their more precise estimates give a better picture of the time path of the work disincentive effects.

HUD's programs of project-based rental assistance, namely, public housing and privately owned subsidized projects, have essentially the same subsidy formula as its housing voucher program. From the tenant's perspective, the subsidy is the difference between the market rent of its unit and 30% of its countable income.²⁰ Some arguments have been offered as to why different types of assistance with the same subsidy formula would have different effects on labor supply. However, the evidence indicates that they have very similar effects.

Jacob and Ludwig and Carlson et al. estimated differences between the work disincentive effects of housing vouchers and public housing and found these differences to be very small. Susin (2005) reached a similar conclusion for the three broad types of assistance using HUD's administrative records to identify the type of assistance received by SIPP households and propensity score matching of assisted with unassisted households to create a sample for analysis. He found that public housing tenants reduce their labor earnings about 17%, tenants in privately owned subsidized projects about 15%, and voucher recipients 14%. Although his sample size is modest (670 assisted households) and his data are not as well suited for the analysis as the data underlying the three preceding studies, the similarity of his results for the voucher program lends credence to his results for the other types of assistance. Olsen et al. (2005) also found small differences in the work disincentive effects of the different types of assistance based on a large random sample of recipients throughout the country (about 150,000 households) from HUD's administrative records for 1995 through 2002 and difference-in-difference estimation. These records contain information on the earnings of each household before it enters the program.

Newman et al. (2009) used methods similar to Susin's methods to match unassisted female-headed households with children in the Panel Study of Income Dynamics (PSID) to similar households living in subsidized housing projects (116 families in public housing and 207 in privately owned subsidized projects). Their point estimates indicate a consistent pattern of work disincentive effects for both types of projects (18 out of 21 for public housing and 17 out of 21 for private projects). They cannot reject the null hypothesis of

²⁰ Although the benefit reduction rate is the same for all, the formula is different for different households because they live in units with different market rents.

no effect at the usual levels of significance in most cases, but the point estimates for the reduction in labor earnings for public housing range from 12% to 29%. The high numbers refer to the early years before much attrition occurs. In those years, the hypothesis of no difference is rejected at the 5% level of significance. For private projects, the effects are much smaller.

Anil et al. (2010) estimated the effect on employment of demolishing public housing projects and offering each occupant a housing voucher or a vacant unit in another public housing project. The projects demolished are usually large projects in poor condition located in neighborhoods with high poverty rates. Unlike MTO, remaining in the current project is not an option. They found that this reform leads to higher employment rates for the former residents of these projects.

Despite being the largest and fastest growing US housing program, there is no evidence on the effects of the LIHTC on labor earnings and employment. Some households served receive subsidies from other programs that limit their rent to 30% of their adjusted income. Presumably, the work disincentive effects for these households would be similar to those in HUD programs with this feature. Other occupants pay a rent that does not depend on their income. The tax credit program almost surely has a much smaller effect on their labor earnings and employment.

The previous studies provide a clear picture of the effects of housing programs on the labor earnings and employment of adults in recipient households. Good evidence on the effects of low-income housing programs on the adult earnings and employment of children in recipient households is scarce. Given the motivations for welfare assistance, this is a serious gap in knowledge.

Based on information in the PSID-Assisted Housing Database about children who were 10–16 years old between 1968 and 1982 and young adults between 1978 and 1993, Newman and Harkness (2002) found that additional years spent in public housing as a child (as opposed to receiving no housing assistance) increased their labor earnings and employment as adults. This evidence does not necessarily argue for public housing or housing assistance in any form. The same amount of money devoted to cash assistance might have had the same effect on earnings and employment. Assistance more targeted on these outcomes would surely have still larger effects (Jacob et al., 2014).

Many people have believed that different housing programs would result in different adult earnings for children in recipient households because they induced recipients to live in different neighborhoods. Indeed, this belief was an important motivation for MTO. Oreopoulos (2003) used data on the random assignment of families to different public housing projects in Toronto to study the effects of neighborhoods on the adult labor market outcomes of the children in these households. He found that neighborhood characteristics have little effect on the adult earnings, unemployment likelihood, or welfare participation of these children.

14.3.4.9 Participation in other welfare programs

Most households eligible for low-income housing assistance in the United States are eligible for other welfare programs such as food stamps and TANF that reduce their benefits with increases in household income. Because they are closely related, many who study the effects of housing programs on labor earnings also study their effect on participation in other welfare programs. The simplest standard model of labor supply helps explain why an offer of housing assistance will make it attractive for some who would otherwise choose not to participate in other welfare programs to reduce their earnings sufficiently to participate in these programs. The offer of housing assistance will also overcome any fixed stigma cost of accepting welfare for some households. The best studies find that low-income housing assistance does induce its recipients to participate in other welfare programs at higher rates. For example, Jacob and Ludwig found that voucher usage increases the TANF participation rate by 15%. [Abt Associates Inc. et al. \(2006\)](#) also found large effects.

A separate question is the effect of housing assistance as a child on welfare program participation as an adult. Standard theory does not have an unambiguous implication about this matter. Based on information in the PSID-Assisted Housing Database on children who were 10–16 years old between 1968 and 1982 and young adults between 1978 and 1993, [Newman and Harkness \(2002\)](#) found that additional years spent in public housing as a child (as opposed to receiving no housing assistance) decreased their participation in welfare programs as an adult.

14.3.4.10 Children's educational outcomes

Older research on low-income housing programs focused primarily on their effects on consumption bundles broadly conceived and the dollar benefit to recipients of changes in them. Recent research has paid increasing attention to consequences of the changed consumption patterns such as educational outcomes of children in assisted households.

Arguably, the primary mechanism through which housing assistance would affect the future labor earnings of the children in recipient households is through its effects on their education. Despite the reduced earnings of adults in the household, housing assistance typically provides their children with better housing and greater consumption of other goods. Both should improve school performance. Housing assistance also provides a better neighborhood for some and a worse one for others with mixed effects on educational outcomes.

The best evidence on the effect of an individual program on educational performance comes from the voucher evaluation. [Abt Associates Inc. et al. \(2006, exhibit 6.5\)](#) found little effect of receipt of a housing voucher (compared with no housing assistance) on education outcomes. The estimated coefficients indicated a mix of small positive and negative effects for its 11 measures. Only one was statistically significant. The results

indicated that children in voucher households were 12 percentage points more likely to have repeated a grade. This might have resulted from transferring to a school with higher standards.

The evidence for subsidized housing projects is not as compelling. Based on information in the PSID-Assisted Housing Database on children who were 10–16 years old between 1968 and 1982 and young adults between 1978 and 1993 and two-stage IV estimation, [Newman and Harkness \(2000\)](#) could not reject the hypothesis that residence in public housing as a child had no effect on the educational attainment at age 20. They found similar results for duration of residence in public housing and the stage of childhood when the child lived there. The results were slightly more positive for occupants of privately owned subsidized projects. Using the data and methods described in discussing their results for housing consumption, [Currie and Yelowitz \(2000\)](#) got somewhat more positive results for public housing. They found that occupancy of public housing in 1992–1993 leads to lower probability of being held back in school, especially for boys and blacks. The Johns Hopkins Longitudinal Study of the Effects of Housing on Health and Social Adjustment (described in more detail in [Section 14.3.4.11](#)) found that a large improvement in housing conditions combined with a modest increase in expenditure on other goods resulting from moving into a new public housing project in the same neighborhood had a tiny effect on the academic achievement of children in the household ([Wilner et al., 1962](#)). Since the improvement in housing was from a base of bad housing by the standards of the 1950s, this strongly suggests that the better housing resulting from housing assistance today would have little effect on educational outcomes for most current recipients.

Finally, [Jacob \(2004\)](#) produced unusually credible estimates of the difference in the effect of public housing and housing vouchers on children's educational outcomes based on data from a natural experiment in Chicago. During the period of the analysis, the Chicago Housing Authority closed some, but not all, buildings in a number of large housing projects in preparation for their redevelopment. Displaced families were offered the option of a housing voucher or a vacant unit in another public housing project. Although some families in other buildings received vouchers through the normal process of reaching the top of the voucher waiting list, displaced families were much more likely to get them. Jacob used occupancy of a building slated for closure as an explanatory variable in regression models explaining the ITT effects on seven educational outcomes of the closures (with their associated alternative subsidies) compared with continuation of the current system, and he used it as an IV in estimating the TOT effect of public housing relative to housing vouchers. The results are based on data from the administrative records of the housing authority and city school system for 10,556 students living in 73 buildings in 9 projects. It includes information for the years before and after the closures were announced. The analysis focuses on children less than 14 years old at the time of the closure announcement. Jacob found that neither the closures nor living in public

housing had much effect on education outcomes. A supplementary analysis suggests why. Neither had much effect on the quality of schools attended as judged by student performance on a standardized math test.

14.3.4.11 Recipient health

The argument that better housing would lead to better health for its occupants and others who come into contact with them was influential in debates over the establishment of low-income housing programs and is still heard today in the United States despite the large improvements in the condition of the unsubsidized housing stock. However, high-quality evidence about the effect of particular low-income housing programs on the health of occupants is rare; good evidence about its effect on others is nonexistent.

A forgotten classic called the Johns Hopkins Longitudinal Study of the Effects of Housing on Health and Social Adjustment provides the best evidence on this matter (Wilner et al., 1962). This study collected many health outcomes for two groups of about 1300 people apiece every 2 months for several years in the mid-1950s. Initially, all were on the waiting list for public housing in Baltimore and lived in the same neighborhood. The control group remained in unsubsidized housing over the period covered. The treatment group moved to a newly built public housing project in the same neighborhood. Each family in the treatment group was matched with a family on the public housing waiting list based on many characteristics including previous health. The public housing project provided substantially better housing (Wilner et al., 1962, pp. 24–26). For example, about 30% of the units occupied by members of the control group did not have bathing facilities in the unit and 25% had rats. The public housing units had neither problem. Beyond an initial period when the children who moved to public housing experienced worse health outcomes, the broad pattern of the results suggests that the substantial improvement in housing conditions led to better health. However, the estimated magnitudes of these effects were usually small, and few were statistically at the 5% level. Since the improvement in housing was from a base of bad housing by the standards of the 1950s, this strongly suggests better housing from current bad housing in developed countries today would have even smaller effects on health outcomes. Few studies have the internal validity of this study. However, it applies to the households that moved into a single low-income housing project in Baltimore.

The best recent study of health effects of programs that subsidize the operation of housing projects is based on data from the Fragile Families and Child Wellbeing Study that collected information between 1998 and 2000 for a random sample of mothers with newly born children. Some, but not all, lived in a subsidized housing project. Using a variety of recalcitrant IV for living in a housing project, Fertig and Reingold (2007) were unable to detect a pattern of significant health effects for this type of housing assistance. For the reasons mentioned in the discussion of Currie and Yelowitz's study, these results surely refer to a mix of public housing and privately owned subsidized projects.

Information about the type of housing assistance received was obtained by asking recipients, and evidence indicates that many are unable to distinguish between the two broad types of project-based assistance.

[Abt Associates Inc. et al. \(2006, exhibit 5.5\)](#) found mixed results for the effect of housing vouchers on its limited self-reported health measures. However, none of the estimated effects are statistically significant at the usual levels.

14.3.4.12 Household composition

The effect of housing assistance on household composition has played little role in housing policy research or debates in the United States. The voucher evaluation produced striking results that might increase interest in this matter. [Abt Associates Inc. et al. \(2006, exhibit 5.3\)](#) discovered that about 18% of the people on the waiting lists for housing vouchers lived with others at some point in the prior year, usually young mothers with children living with their parents or other relatives. When offered assistance, they chose to live independently. Voucher receipt reduced the number of families that shared housing with others by about 22 percentage points. It also reduced the mean size of the household by almost three quarters of a person and the number of intergenerational households by 20 percentage points ([Abt Associates Inc. et al., 2006, exhibit 3.10](#)). [Ellen and O’Flaherty \(2007\)](#) produced similar results for public housing as well as the housing voucher program.

14.3.4.13 Effects on neighbors of assisted households

The impact of low-income housing projects and families with housing vouchers on neighborhoods has been a perennial issue in popular debates over housing policy. The popular view is that they make their neighborhoods worse places to live, primarily because they bring less desirable people into the neighborhood. However, there are other reasons to expect subsidized projects or housing vouchers to improve neighborhoods in certain circumstances ([Ellen et al., 2007](#), pp. 263–264). A new subsidized project built at low-density in a neighborhood with the worst housing and poorest families would almost surely make that neighborhood a more attractive place to live for many years after its construction.

If a housing project makes its neighborhood a better or worse place to live, this would be reflected in neighborhood property values. Early studies of this matter were limited to a small number of projects in a single city or based on crude methods and data.²¹ With one major exception, recent studies have also been limited to a single city or metro area. However, the best have assembled impressive data sets and used superior methodologies to analyze them. As a result, we have much better information about the effects of different housing programs on the neighborhoods of their participants.

²¹ [Galster et al. \(1999a, chapter 4\)](#) provided a review of this literature.

In a precursor to the recent studies, [Lee et al. \(1999\)](#) studied the effects of all major urban rental housing programs in Philadelphia. Regressing the sales price of single-family units on the number of assisted rental units of each type within a fourth and an eighth of a mile and demographic, housing, and amenity variables, they found small positive effects on neighboring property values on average for some programs and small negative effects for others.²²

[Galster et al. \(1999b\)](#) refined Lee, Culhane, and Wachter's methodology in a study of the effect of housing voucher recipients on neighborhood property values. The voucher program serves primarily the very poorest households (more than 75% have incomes less than 30% of the local median), and it creates strong incentives to occupy a dwelling unit with about median rent. Therefore, voucher recipients could move into neighborhoods populated largely by families with much higher incomes. However, the evidence indicates that they don't typically do that. Because most voucher recipients move to neighborhoods that are only slightly better than their initial neighborhoods, their effect on neighborhood property values is likely to be modest in most cases.

Galster, Tatian, and Smith's study uses panel data from Baltimore in the mid-1990s on the sales prices of single-family units, their characteristics, and the number of rental buildings and units within certain distances of them that house voucher recipients. To account for the difference between the unobserved characteristics of the neighborhoods of single-family units that affect both sales prices of houses and the number of voucher recipients who end up in the neighborhood, the regression model distinguishes between neighborhoods that do and do not attract voucher recipients at some time during the period studied. They found statistically significant effects of the occupancy of units by recipients of housing vouchers on neighborhood property values. However, the direction of the effect depends on the nature of the neighborhood and the concentration of program participants in it, and the magnitudes are relatively small in most cases. Because they used sample selection criteria that ruled out more than 98% of the buildings with voucher recipients, it is not clear whether these results generalize to the entire voucher program in Baltimore at that time. The regression models also involved restrictive assumptions about the changes in property values over time that have been relaxed by others in later work.

Using the same methodology, [Santiago et al. \(2001\)](#) found a modest positive effect on neighborhood property values of Denver's dispersed public housing program in most cases. Under this program, the Denver Housing Authority purchased single-family houses, duplexes, and small multifamily buildings throughout the city and rented them to people on the public housing waiting list. In assessing the results, it is important to

²² They incorrectly assumed that the Federal Housing Administration (FHA)-assisted units in their data are owner-occupied units. In fact, these are units in privately owned subsidized rental projects under programs such as Section 221(d)(3) and 236 whose mortgages are insured by the FHA. Therefore, their results don't pertain to homeownership programs.

recognize the program's unique features. These sites were required to be highly dispersed, their occupants were selected for their good behavior, and the housing authority followed a strict inspection and maintenance policy for these units.

In a well-designed study that contains important refinements of Galster, Tatian, and Smith's methodology and is based on an impressive data set for New York City, [Schwartz et al. \(2006\)](#) found substantial positive effects on neighborhood property values of a collection of subsidized construction and rehabilitation programs that constituted the city's ambitious 10-Year Plan for neighborhood revitalization of its most distressed areas. This result arguably illustrates the importance of the location of subsidized projects. In a similar study with New York City data, [Ellen et al. \(2007\)](#) found that the major federal housing construction programs have not typically led to reductions in neighborhood property values and in some cases increased them. Using the same data set, [Ellen and Voicu \(2006\)](#) found that projects developed by nonprofit organizations have a more positive effect on neighborhoods than those developed by for-profits.

One concern about these results is that they apply only to New York City. This city has one of the country's most regulated housing markets. For example, it has a moderate form of rent control. [Baum-Snow and Marion \(2009\)](#) provided some evidence that the results apply to other parts of the country. They used RD methods and national data to study a variety of effects of LIHTC projects on neighborhoods. Like [Ellen et al. \(2007\)](#), they found that the construction of subsidized housing projects increases neighborhood property values in declining neighborhoods. In stable and gentrifying areas, they found no effect.

To understand why subsidized housing projects have not usually reduced neighborhood property values, it is important to recognize the ability of local political bodies to influence their location. For public housing, they have had veto over location. For privately owned subsidized projects, they could exert considerable influence through local land use regulations, for example, through granting variances from zoning ordinances. Due to neighborhood opposition, projects are rarely built in the best neighborhoods or indeed in any neighborhood with more attractive housing. It is also important to recognize the incentives facing developers of privately owned subsidized projects. For example, the largest program (LIHTC) provides a much larger subsidy to projects in census tracts where more than half of the households have incomes less than 60% of the local median.

[Rossi-Hansberg et al. \(2010\)](#) recently developed a novel approach for estimating the effects of housing improvements on nearby land prices and applied it to a unique program in Richmond, Virginia, that provided concentrated housing assistance in a few small areas. The funds to support this program came primarily from HUD block grants to state and local governments, namely, the Community Development Block Grant and the HOME Program. These programs permit a wide variety of different types of housing assistance, for example, loans at below-market interest rates to owner-occupants to

rehabilitate their houses and subsidies to selected developers to build new units to rent or sell to low-income households. The authors did not attempt to estimate the effects of different types of assistance. Indeed, they didn't report the mix of types used in the program studied. Although the paper combines high-quality theoretical and empirical analysis, the estimated effects of the program seem implausibly large. In three of the four sites, the estimated increase in land prices within 3500 feet of the heavily subsidized areas was five to seven times larger than the amount of the housing subsidies. In the other site, it was twice as large. The authors recognize several possible culprits and explore them to the extent possible with the readily available data. The leading candidate is that nearby structures were improved in ways that could not be detected with their data and that this might have led to the large estimated effects on land prices. The improved appearance of the neighborhood resulting from the concentrated housing subsidies might have induced families with higher incomes to move into the areas and these people might have demanded better housing.

The methodology used in all studies is based on the assumption that subsidized housing projects have no effect on property values in other neighborhoods. In essence, the authors obtain their results by comparing changes in property values in the neighborhoods of subsidized projects with the changes that occur in other initially similar neighborhoods. A positive estimated coefficient of the variable in the hedonic equation indicating that the property sold is in the neighborhood of a subsidized housing project leads the authors to conclude that housing projects have made their neighborhoods better places to live. The problem with this conclusion is that the authors would have gotten the same estimated coefficient if these housing projects had no effect on neighboring property values and reduced property values in other neighborhoods by the magnitude of this coefficient. In this case, the usual interpretation of the results would be completely reversed. Obviously, there are infinitely many other combinations of changes in property values in the neighborhoods of subsidized projects and other neighborhoods that would yield the same coefficient in the hedonic regression equation. What this coefficient really identifies is the difference between the effects of subsidized housing projects on property values in the two types of neighborhoods. To get their conclusion from their evidence, the authors implicitly assume that subsidized housing projects have no effect on property values in neighborhoods without these projects.

There are several reasons to object to this assumption. First, some households that moved into the subsidized housing projects previously lived in other neighborhoods, and their moves decreased the demand for housing in these neighborhoods. Second, some households that lived in the neighborhoods of the subsidized housing projects right before they were built would be forced to move because their units were torn down in order to build a new project. Others from the neighborhood that didn't move into the subsidized project would want to move to another neighborhood if the subsidized housing project makes their initial neighborhood a sufficiently better or worse place to live. In

equilibrium before the construction of a subsidized housing project in their neighborhood, each household has chosen a dwelling unit and neighborhood with the utility maximizing combination of characteristics given the household's preferences, income, and the equilibrium prices of housing units with each combination of characteristics. If subsidized housing projects change the desirability of neighborhoods, initial residents will no longer be in equilibrium at the initial prices. The movement of these households will lead to changes in property values in other neighborhoods. [Baum-Snow and Marion \(2009\)](#) produced results consistent with this reasoning. They found that LIHTC projects increase neighborhood turnover among homeowners.

One important issue that hasn't been studied is who gains and who loses if a subsidized housing project improves a neighborhood. A popular view is that it is the people living in the neighborhood when the project is built. However, economic theory suggests that the primary beneficiaries of neighborhood upgrading are the owners of nearby properties. As a result of zoning laws, much of the housing surrounding subsidized housing projects is surely rental. If a newly built subsidized project makes the neighborhood a more attractive place to live, the owners of this rental housing will charge higher rents and the value of their property will be greater. Since the occupants of this rental housing could have lived in a nicer neighborhood prior to the project by paying a higher rent, they are hurt by its construction. The poor in the project's neighborhood will benefit from the neighborhood upgrading only to the extent that they own the property surrounding the project or move into the project.

Low-income housing programs affect the desirability of neighborhoods in part by changing the people who live in them. One of the most contentious issues in popular debates over housing policy is the extent to which the movement of subsidized households into a neighborhood increases crimes against its initial residents. This is one mechanism through which housing programs could affect neighborhood property values. Several recent articles shed light on this matter.

Using a panel of annual data from 1996 to 2008 at the census tract level for 10 large cities, [Ellen et al. \(2012\)](#) found a negligible effect of an additional voucher recipient in a neighborhood on the total number of crimes as well as property and violent crimes separately. Because most voucher recipients move to neighborhoods that are only slightly better than their initial neighborhoods, this is not surprising. The results also indicate that an additional unit in a public housing project has a similar effect as an additional voucher recipient. Because the mean number of households in a public housing project is about four times the mean number of voucher recipients in a census tract, a single public housing project in a census tract would typically have a much larger effect on the crime rate in that census tract than the housing voucher program. In the 10 cities involved in this study, the census tract with the most public housing units served about 6500 families in public housing projects. The census tract with the most voucher recipients served about 350 recipients.

Stimulated by a magazine article that received considerable attention in the popular press, Popkin et al. (2012) focused on the effects of voucher recipients who were displaced from public housing on violent and property crime rates in their destination neighborhoods. When their buildings were demolished or substantially rehabilitated, public housing tenants are offered the option of a housing voucher or a vacant unit in another public housing project. The authors also consider the effects of other voucher recipients on crime rates. This study used quarterly census tract data over 8–10 years in Atlanta and Chicago, the two cities with the largest public housing transformation efforts. Both demolished most of their public housing units. With a lag, the land was typically used to build new, often mixed-income, housing at lower density. Popkin et al. used separate methods for estimating the effects on crime in the census tracts of the demolished public housing units and other census tracts. Like Ellen, Lens, and O'Regan, they found that regular voucher recipients have negligible effects on crime rates in their destination neighborhoods. Voucher recipients who are forced to relocate from public housing projects have positive effects on neighborhood crime rates that are an order of magnitude larger. Their evidence indicates that the crime rate increases at an increasing rate with the number of relocated voucher recipients moving into the census tract. In the census tracts where public housing projects were redeveloped, crime rates fell markedly. The net effect was a small overall reduction in crime rates.

Freedman and Owens (2011) studied the effect of LIHTC projects built in qualified census tracts on the major types of violent and property crimes in their counties.²³ As mentioned earlier, qualified census tracts are areas with unusually large numbers of the poorest families where tax credit developers receive substantially greater subsidies. Using a panel of national data from 2000 to 2007 and two-stage least squares estimation with the fraction of the county population in a qualified census tract as an instrument for the number of tax credit units in qualified census tracts, they found that additional tax credit units in these areas reduce robberies and increase motor vehicle thefts. The effects on other types of crime are statistically insignificant at the 5% level. Since they found that tax credit development in these census tracts is largely at the expense of tax credit development elsewhere in the county, these are mainly the effects of the location of the projects. The magnitudes of the effects are substantial. For example, an additional tax credit project of mean size (about 75 units) located in a qualified census tract rather than elsewhere in its county is estimated to reduce the county robbery rate by 10.5% and increase motor vehicle thefts by 4.7%.²⁴ However, based on their results and estimates of the effect of police spending on crime from other studies, they concluded that the same reduction in robberies could be achieved by spending on police a small fraction of the additional subsidy provided to tax credit developments in qualified census tracts.

²³ In the United States, the average county has 24 census tracts.

²⁴ Our thanks to the authors for providing these additional results.

To understand the effects of tax credit projects on their neighborhoods, it is important to keep two things in mind. First, tax credit projects often involve tearing down residential structures of very low quality occupied by families with very low incomes. Some of the displaced families surely move to another county. Second, tax credit projects that are not renovations of subsidized projects built under other low-income housing programs serve families with incomes about twice as high as the families served by the other programs (GAO, 1997, p. 146). Therefore, many tax credit projects surely increase neighborhood income.

14.3.4.14 Market prices

The belief that housing vouchers would lead to higher rents for unsubsidized low-income households has been influential in housing policy debates since the debate over the creation of the public housing program in the 1930s. To resolve disagreements about this matter, the US Congress authorized the largest social experiment in the country's history. EHAP's Housing Assistance Supply Experiment (HASE) operated entitlement housing allowance programs in the Green Bay and South Bend metropolitan areas. This study found little effect of housing allowances on the market rents of units of any type (Rydell et al., 1982; Lowry, 1983). For units that were significantly below standards prior to the experiment, rents fell slightly. For modest units meeting the standards or falling slightly below them, rents rose slightly. A careful reassessment of the HASE evidence reached the same conclusion (Mills and Sullivan, 1981). In a paper that calibrates the Anas and Arnott general equilibrium model of the housing market with data from Los Angeles, San Diego, San Francisco, and Sacramento, Mansur et al. (2002, table 4) got similar results for the effects of an entitlement program that would provide a cash subsidy to all renters in the lowest income quintile.

If an entitlement housing allowance program for which 20% of households were eligible had no discernible effect on housing prices, it is reasonable to conclude that existing tenant-based programs have little effect. A more recent study concludes that the housing voucher program has increased rents of low-quality housing by an average of 16% (Susin, 2002). This conclusion is inconsistent with the HASE results and the clear implications of standard economic theory.²⁵ In the absence of vouchers, their recipients would occupy the least desirable housing. Vouchers enable these households to occupy rental units of average quality. Hence, additional vouchers will decrease the demand for low-quality housing and increase the demand for average quality rental housing. If vouchers affect the price of rental housing, their introduction should, in the short run, decrease the price of low-quality housing and increase the price of average quality units. Using data from the same source as Susin (the American Housing Survey) but a different methodology, Eriksen and Ross (forthcoming) got results that accord with these theoretical

²⁵ Olsen (2003, pp. 421–422) offered a possible explanation for Susin's results.

expectations. In the short run, additional housing vouchers drive down market rents of the least desirable units and increase the rents of units of about average desirability by modest amounts. These price effects are somewhat greater in markets with the least elastic housing supply.

The belief that subsidized construction programs would lead to lower rents for unsubsidized low-income households has also been influential in US housing policy debates. [Olsen \(2007b, p. 622\)](#) explains why this belief is inconsistent with the standard assumptions of economic theory in the long run and not an implication of these assumptions in the short run. In the simplest economic model of competitive markets, all firms that supply a good are identical in the long run and the long-run equilibrium market price is equal to the minimum long-run average cost of production. If subsidized construction programs increase housing output and decrease output of other goods when account is taken of both the subsidies and taxes needed to fund them, they lead to increases in the prices of inputs that are used most heavily in the production of housing relative to other goods and decreases of the prices of inputs used least heavily in the production of housing. The net effect of these changes in input prices is to increase the minimum long-run average cost of production of housing services and hence its long-run equilibrium market price. This is the standard explanation for an upward sloping long-run supply curve for a good. So subsidized construction programs lead to an increase in the long-run equilibrium price of housing services facing unsubsidized households. In the short run, subsidized construction programs decrease both demand and supply in the unsubsidized housing market. This gives us no reason to believe that these programs decrease the market price of housing services. However, no one has attempted to produce evidence about this matter.

A comprehensive analysis of the effects of the voucher program on market prices would consider not only housing prices but also the prices of other goods. If a housing voucher program (or indeed any housing program) does increase the consumption of housing services by recipients more than it decreases housing consumption by taxpayers thereby increasing total housing consumption, it must reduce the demand for nonhousing goods and hence their market prices unless the supply curves for all other goods are perfectly elastic. Taxing some people to provide housing assistance to others does not enable the economy to produce more of all goods. The changes in market prices of housing and other goods will benefit some consumers and hurt others. The change in the pattern of output will also lead to increased demand for some inputs and decreased demand for other inputs with resulting changes in input prices and hence gains to some and losses to others. To the best of our knowledge, these effects of housing programs have never been studied.

14.3.4.15 Crowding out of private provision

Many people with great influence on housing policy hold the simplistic view that subsidized construction programs add one unit to the housing stock for each unit built and

housing vouchers have no effect on the size of the housing stock, and these views are influential in leading them to support a very cost-ineffective approach to providing housing assistance. Murray (1983, 1999), Malpezzi and Vandell (2002), Sinai and Waldfoegel (2005), and Eriksen and Rosenthal (2010) showed how far this simplistic view is from reality. They found very substantial crowding out of private provision by construction of subsidized projects, well above 50% in almost all studies and close to 100% in others. The results differ somewhat across studies in part because some deal only with the crowd out of unsubsidized new construction and others account for withdrawals of existing units from the housing stock.

A more sophisticated view is that new construction programs will increase the number of housing units, albeit not one for one, and that tenant-based housing vouchers will have little effect on the housing stock. However, as Murray (1999) had pointed out, all subsidized housing programs lead to an increase in the number of dwelling units by increasing the demand for distinct units. The offer of housing assistance of any type induces individuals and families living with others to live in their own units. Abt Associates Inc. et al. (2006, pp. 23, 76) indicated that about 26% of the families on the housing voucher waiting list were living with friends or relatives and 2% were living in a homeless shelter or transitional housing, and voucher usage resulted in corresponding decreases in these numbers. Since doubling up and homelessness are more common among the poorest households, the programs that serve the poorest households will have the greatest net effect on the number of housing units. The voucher program serves somewhat poorer households than public housing and much poorer households than privately owned subsidized projects as judged by *per capita* household income (Olsen, 2003, p. 393). Consistent with this explanation, Sinai and Waldfoegel (2005) found that tenant-based vouchers lead to a larger increase in the housing stock than construction programs. This same phenomenon could explain the difference that Murray (1983, 1999) found in the extent to which public housing and privately owned projects crowd out unsubsidized construction. At the time of his data, public housing served much poorer households than private projects (Olsen and Ludwig, 2013, table 8.2). It also explains Eriksen and Rosenthal's finding of almost complete crowd out for the LIHTC. This program serves families with much higher incomes than the other programs.

14.3.4.16 Miscellaneous

Other research that does not deal directly with the performance of particular low-income housing programs provides important insights for housing policy development. The MTO research that focuses generally on the effects of living in a low-poverty neighborhood is in this category. It is relevant for decisions about the extent to which housing policies should promote economic integration. Another recent example is research confirming that the filtering process works well to supply housing to low-income households

(Rosenthal, 2014). This is relevant for decisions about the extent to which housing policies should rely on housing vouchers versus subsidized construction. A few studies such as Early and Olsen (2012, section 4) have estimated important effects of fundamental reforms of the current system designed to eliminate its substantial inefficiencies, inequities, and bias against homeownership.

14.3.5 Conclusion

The existing evidence on the performance of low-income rental housing programs in the United States suggests a number of conclusions. All major housing programs have increased substantially aggregate housing consumption by recipients and almost all have increased significantly their consumption of other goods. Voucher recipients have lived in slightly better neighborhoods as a result of the program, occupants of privately owned subsidized projects in slightly worse neighborhoods, and occupants of public housing projects in noticeably worse neighborhoods. All programs have increased aggregate housing consumption more than would have occurred if each participant had been given an unrestricted cash grant equal to his or her housing subsidy. The net effect of these changes in consumption patterns is that housing programs have typically provided large benefits to their recipients. Although mean benefit has been large compared with mean recipient income, it has been small compared with the cost to taxpayers. The mean benefit has been less than 75% of the mean subsidy for programs that have subsidized the construction of housing projects and about 80% for vouchers. For vouchers, the cost to taxpayers has exceeded the subsidy by the modest administrative cost. For housing projects, the cost to taxpayers has been much larger than the sum of the subsidy and administrative cost. The mean benefit of each program studied varies inversely with income and directly with family size, but the variance in real benefits among similar households is large under most programs.

US low-income rental housing programs have induced adult recipients to earn 10–15% less on average, and they have induced recipients to participate in other welfare programs at higher rates. They have had little effect on the health of recipients or the educational outcomes of the children in these households. They have induced many families that would have lived with others in the absence of housing assistance to live independently.

Most existing studies find small positive effects on neighbor property values of some housing programs and small negative effects for others. The magnitude and direction of these effects depend on the type of neighborhood involved. Housing programs have had small effects on the rents of unsubsidized units that are not located near subsidized units.

The most important finding of the empirical literature from the viewpoint of housing policy is that tenant-based housing vouchers have provided equally good housing at a much lower cost than any type of project-based assistance that has been studied. This

finding implies that a shift of all discretionary resources from programs of project-based assistance to tenant-based vouchers would enable us to provide several million additional households with adequate housing at an affordable rent without any increase in government expenditure.

Although important progress has been made over the past 15 years in understanding certain aspects of the performance of low-income housing programs, this review of the literature reveals many gaps and weaknesses in our knowledge.

For several large programs, there is little or no evidence on most aspects of performance. Evidence on the programs that have accounted for the bulk of additional housing assistance over the past 15 years is especially meager. There is no credible evidence about the performance of the HOME Investment Partnerships Program, a housing block grant to state and local governments that has existed for more than two decades and spends almost \$2 billion a year. Due to the dearth of easily accessible data on the characteristics of the occupants of tax credit units and the nature of their housing and the abundance of data on the number of units built and their location, the limited research on the tax credit program has focused on a few aspects of its performance, primarily on its effects on the neighborhoods in which the projects are located and the extent to which it has crowded out private construction.

This survey also reveals a dearth of recent evidence on the performance of programs that no longer subsidize the construction of housing projects but continue to subsidize their operation. These programs account for about 40% of the budget for low-income rental housing assistance and serve a similar fraction of all recipients of this assistance. Evidence on their performance is important for making good decisions about discretionary expenditures on these programs.

There is no evidence about some potentially important effects of housing programs such as the consumption distortions that result from the method used to ration dwelling units in housing projects, namely, offering the family that reaches the top of the waiting list an all-or-nothing choice of a particular unit. Evidence on many other outcomes such as the health of recipients and the educational achievements of children in these households is sparse. Evidence on still other important outcomes is old and suffers from selection bias. As a result, we don't have good answers to key questions such as whether current housing programs lead to greater overall consumption of housing services than would result from unrestricted cash grants in amounts equal to the housing subsidy or taxpayer cost.

There is no recent high-quality evidence of the cost-effectiveness of different methods of delivering housing assistance such as discretionary expenditures on the LIHTC and HOME programs, renewals of use agreements with owners of privately owned subsidized projects, and public housing operating and modernization subsidies to guide the allocation of funds across programs. Evidence on this matter is particularly important because previous research suggests that cost-effectiveness in providing housing is the largest and most consequential difference in the performance of different housing programs.

Because no studies in the past two decades and few studies before then compare the consumption patterns of recipients of housing assistance with what their consumption patterns would have been had they been given equally costly unrestricted cash grants, it is impossible to say whether current housing programs are consistent with their rationales. Do they induce recipients with the weakest taste for housing relative to other goods to occupy better units than would result from cash grants? Do they induce recipients with the strongest taste for housing relative to other goods to occupy worse units? Housing programs must have these outcomes in order to be consistent with their primary rationales. Our knowledge of the tangible external effects of low-income housing programs also leaves much to be desired. For example, what housing features generate the tangible external benefits? Without the answer to this question, it is impossible to design a cost-effective housing subsidy to address the external effect.

Although governments in many other countries play a much larger role in providing housing assistance to low-income households, the English-language literature about the performance of housing programs outside the United States is extremely modest.²⁶ It is possible that there is a plethora of excellent studies in the native languages of these countries, but we are skeptical. One problem that has dogged scholars studying the effects of low-income housing programs outside the United States is a dearth of data on detailed characteristics of housing units and their neighborhoods. It appears that other countries do not have a data set nearly as detailed as the American Housing Survey in these regards. This has made it difficult to disentangle the effects of the programs on the price as opposed to the quantity of housing services. Since different programs in the same broad category such as public housing can have very different structures in different countries, they might have very different outcomes. Indeed, due to differences in regulations of private markets across different countries, it is reasonable to expect some differences in the effects of identical programs across countries. For these reasons, research on the performance of the large housing programs outside the United States would contribute greatly to housing policy development in these countries.

The available empirical evidence has much to contribute to the design of housing policy reforms. However, the magnitude of the public expenditures involved argues for producing better information on which to base these decisions. Evidence based on recent data on the full range of major housing programs for low-income households is badly needed.

²⁶ The UN-Habitat's Adequate Housing Series (<http://unhabitat.org/series/adequate-housing-series/>) describes housing policies throughout the world. Whitehead and Scanlon (2007) provided a more detailed account for European countries. Bingley and Walker (2001), Gibbons and Manning (2006), Hills (1991), Laferrere and Le Blanc (2004), Le Blanc and Laferrere (2001), Lui and Suen (2011), and Wong and Liu (1988) analyzed some effects of a few of these programs.

14.4. US HOMEOWNERSHIP POLICY

US housing policy can affect the homeownership rate directly through the tax system and indirectly through the mortgage market. The former includes the tax exemption for mortgage interest and local property taxes, the nontaxation of homeowners' imputed rental income, and tax-exempt bond financing (such as mortgage revenue bonds). Given that these topics have been covered extensively in the literature, we will focus on the impact US policy has had on the mortgage market and how this translates into changes in the homeownership rate.²⁷

Many political leaders have argued for government policies to promote homeownership on the grounds that homeownership is the "American Dream." As Shaun Donovan, Secretary of HUD, recently put it in a White House blog, "owning a home was a symbol of responsibility and a source of security for millions of middle-class families across the country."²⁸ As such, homeownership has been actively promoted by the federal government, first by Presidents Hoover and FDR but most recently by Presidents Reagan, Carter, Clinton, George W. Bush, and Obama.²⁹ HUD's goal was to reach a 70% homeownership rate by 2006 (in fact it reached 69% in 2004).³⁰ Federal housing policy is based on the premise that the homeownership rate that results from market forces alone is sub-optimal. We discuss justifications for promoting homeownership in [Section 14.4.1](#).

US housing policy that is related to the mortgage market began as a direct response to the severe problems in the housing market brought on by the Great Depression. This is where our analysis begins and continues up to the recent financial and housing crisis. To provide context for our analysis, we present data on the US homeownership rate starting in 1890 in [Section 14.4.2](#) and then a brief history of the role of the federal government in the mortgage market in [Section 14.4.3](#).

We next address the issue of whether there is any causal evidence linking federal intervention in the mortgage market to homeownership rates in [Section 14.4.4](#). This includes an analysis of whether the affordable housing goals as established in the CRA in 1977 and

²⁷ Hilber and Turner (2014) concluded that the existing literature finds little impact of the mortgage interest rate deduction on the homeownership rate, though they show that a positive impact occurs for higher-income households in MSAs with low levels of land use restrictions. Durning and Quigley (1985) and Benjamin and Sirmans (1987) showed that the subsidies provided to homebuyers through mortgage revenue bonds are capitalized into house prices and hence are not likely to have a significant impact on the homeownership rate.

²⁸ <http://www.whitehouse.gov/blog/2013/08/06/promoting-american-dream-homeownership>.

²⁹ Hardaway (2009, p. 46) says, "Although there appears to be a common perception that it was President Roosevelt's New Deal that first fostered the notion of home ownership as a national ideal, it was, in fact, Herbert Hoover who in 1932 pushed through the Federal Home Loan Bank Act of 1932, with the purpose of providing liquidity to mortgage lenders, creating a secondary market, and 'promoting home ownership.'"

³⁰ See Gabriel (2001).

in the GSE Act of 1992 had any role in the recent financial crisis. One can think of a foreclosure as a loss of homeownership so we look at the role of federal and state policies that relate to foreclosure rates. We then analyze the effectiveness of the policies that the federal government enacted to ease the severity of the recent housing market/foreclosure crisis in [Section 14.4.5](#). We end this section with a summary of our assessment of the empirical literature that evaluates state and federal interventions in the mortgage market ([Section 14.4.6](#)).

14.4.1 Justifications for promoting homeownership

Relatively speaking, US homeownership policy has not targeted low-income households so its justifications are different from those for low-income rental assistance. We now describe five rationales for homeownership that have appeared in the literature. First, positive externalities can justify policies that promote homeownership. These externalities include increased exterior maintenance, household and neighborhood stability, becoming more “vested” in maintaining local public goods, more civic engagement, and general increases in social capital ([Green and White, 1997](#); [DiPasquale and Glaeser, 1999](#); [Coulson et al., 2003](#); [Dietz and Haurin, 2003](#); [Hilber and Mayer, 2009](#); [Coulson and Li, 2013](#)). It has been argued that neighborhood improvement that results from homeownership is particularly valuable in low-income areas ([Shlay, 2006](#)). There is also some evidence that homeownership also benefits children and these benefits are typically not fully internalized into homeownership decisions ([Green and White, 1997](#); [Boehm and Schlottmann, 1999](#); [Harkness and Newman, 2003](#)).

Second, another externality relates to the lack of information in thin markets. The argument is that originators base their decision to accept or reject a mortgage application based on the appraisal value of the home. Since appraisal values are based on comparables, the more transactions that occur nearby, the more accurate will be the appraisal. If the local market is thin, as it can be for low-income neighborhoods, appraisals will be measured with more error and this increased risk can lead originators to reject some marginal loans. By promoting transactions in the underserved areas, government action can help to overcome this lack of information ([Harrison et al., 2002](#); [Voicu et al., 2014](#)).

Third, imperfections in the mortgage market can motivate government involvement. In particular, information asymmetries between borrowers and lenders can result in the classic lemons problem where only risky borrowers buy loans at interest rates that are too high for low-risk borrowers to participate in the market. This can justify the existence of government agencies that promote liquidity in the primary mortgage market. This seems to be particularly relevant during economic crises such as the Great Depression and the Great Recession. A second source of information asymmetry exists between mortgage originators and purchasers in the secondary market where originators have more knowledge about the quality of loans and have an incentive to sell the riskier loans. Again, this

can result in illiquidity in the secondary market. Government involvement can promote liquidity in the secondary market, particularly during times of great financial distress (Glaeser, 2010).

Fourth, discrimination in the mortgage market can exist. This can result in credit not being extended to creditworthy borrowers because of unrelated characteristics such as race. As noted by Avery et al. (2005), extending credit to these underserved groups should result in an increase in profitable lending.

Fifth, for homeowners, housing is their largest asset. So house price appreciation is an important means for increasing household asset holdings, particularly for low-income households (Belsky and Duda, 2002; Boehm and Schlottmann, 2008; Herbert et al., 2014). Furthermore, paying down the mortgage principal can be seen as a savings mechanism that adds to home equity. Generally, this asset accumulation and savings are viewed as a positive outcome for the economy. Given that households do not fully internalize the value of homeownership as a means for asset accumulation, there is a role for government involvement.

For the most part, the net benefits of addressing these market imperfections have not been accurately measured, so whether federal involvement in the housing market has improved the situation from an economic standpoint is largely an open question. As previously mentioned, a primary and longstanding federal justification for promotion of homeownership is that it is the “American Dream.” Based on this discussion, it is not clear that the federal interventions motivated by this belief have been welfare improving. Bearing this in mind, we focus on a thorough investigation of the empirical work examining the impact of federal intervention in the mortgage market on homeownership rates and the government response to the housing crisis in Sections 14.4.4 and 14.4.5.

14.4.2 The homeownership rate in the United States

The homeownership rate starting in 1890 is plotted in Figure 14.1. Figure 14.2 plots the annual rate starting in 1965 when information is available on a yearly basis. The homeownership rate was fairly steady (46–49%) between 1890 and 1930. It then fell by 5 percentage points as a result of the Great Depression and bottomed out at 43.6% in 1940. It then grew by an astronomical 11.4 percentage points over the next decade and by 6.9 percentage points between 1950 and 1960 to a rate of 61.9%. A fact that has gone unrecognized in much of the recent economics literature is that more than half of the overall increase over these two decades took place by the end of 1945 (Fetter, 2013b). The homeownership rate increased by only 1 percentage point between 1965 (63.0%) and 1994 (64.0%). It then steadily increased and peaked in 2004 at a rate of 69.0%. The ensuing housing crisis has resulted in a slow and steady decline to a rate of 65.1% in 2013, which erased most of the gain made in the previous 10 years.

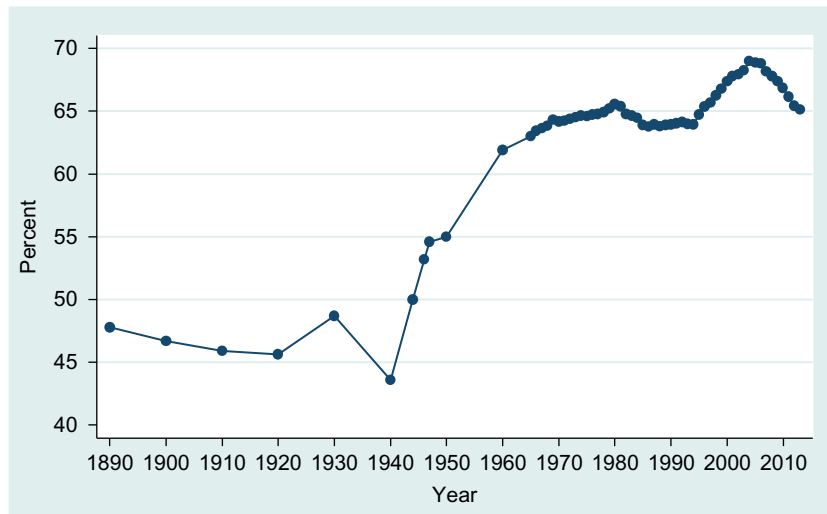


Figure 14.1 US homeownership rate: 1890–2012. Source: *National Historical Geographic Information System, US Census Bureau, and Fetter (2013a,b).*

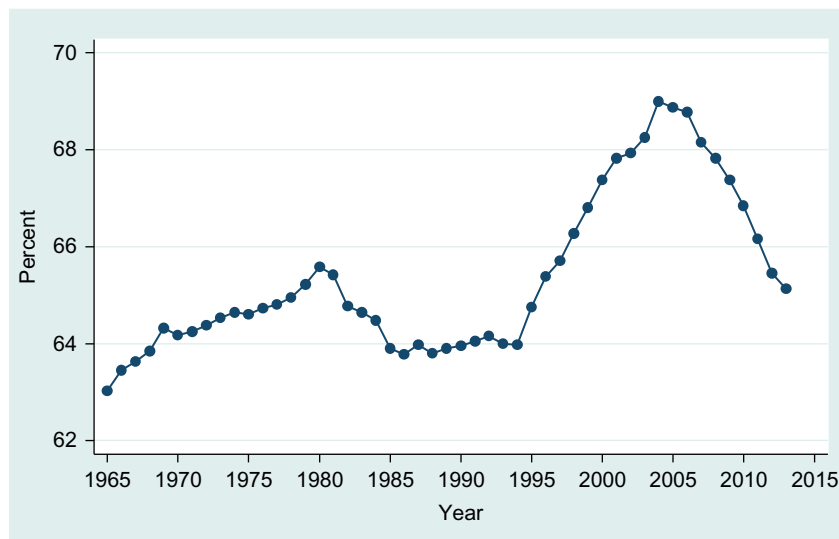


Figure 14.2 US homeownership rate: 1965–2013. Source: *US Census Bureau Housing Vacancy Survey.*

The black homeownership rate is estimated to be 24.6% in 1930 and 22.8% in 1940. The homeownership rate by race and ethnicity is given in [Figure 14.3](#) for 1994–2013. There was actually a decline in the black homeownership rate between 1980 and 1995 before it rose by about 7 percentage points over the next 9 years, ending at its highest level

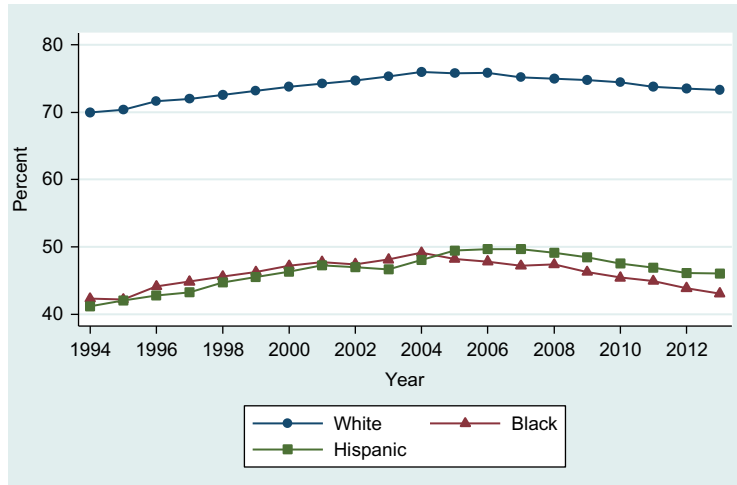


Figure 14.3 US homeownership rate by race/ethnicity: 1994–2013. Source: US Census Bureau Housing Vacancy Survey.

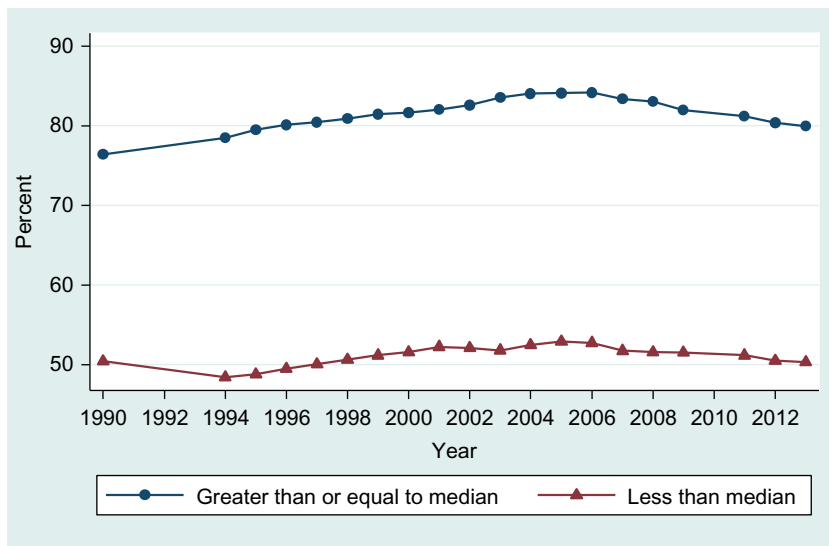


Figure 14.4 US homeownership rate by income: 1990–2013. Source: US Census Bureau Housing Vacancy Survey and NHGIS.

ever at 49.1% in 2004. It then declined by 6 percentage points to a level similar to what it was in 1990 but lower by about 1 percentage point than the level in 1980.

Figure 14.4 plots the homeownership rate by two income groups: those with family incomes below the median and those with family incomes greater than or equal to the

median for 1990–2013. What is clear is that any attempt to increase homeownership for the lower half of the population in terms of family income has not been successful. First, note that the rate actually decreased from 50.4% in 1990 to 48.4% in 1994. Second, while the homeownership rate for the lower half did increase to 52.9% in 2005, it has fallen back to 50.3% in 2013—almost identical to the rate in 1990.

The question to answer is what role did the federal government play in determining the US homeownership rate? There are two means through which the federal government has promoted homeownership: the tax (and subsidy) system and the regulation, support, and participation in the mortgage market. We will focus on the latter, particularly as it relates to the recent financial crisis.

14.4.3 The history of the federal role in the mortgage market

The federal government involvement in the mortgage market has been a complex story that involves a dizzying array of agencies, programs, laws, and policies (and acronyms!). It has witnessed the evolution of the mortgage instrument from a short-term, nonamortizing, balloon payment product with low loan-to-value (LTV) ratios to the very complicated array of choices that exist today. This occurred during and potentially affected wide swings in the housing market including the recent financial/housing crisis. And as just discussed, this has all happened when there have been large changes in the homeownership rate in the United States. In order to fully understand the causal link between policies that are related to the mortgage market and the homeownership rate, it is crucial to provide a brief history of the federal role in the mortgage market. For example, to understand what part the GSEs may have had in the recent financial crisis and how this may have affected the homeownership rate, it is useful to have knowledge of their genesis and evolving role in the mortgage market.

This brief summary of the history of the federal involvement in the mortgage market starts in the early 1930s and covers through the recent financial crisis. We will then take up the government response to the financial crisis and its effectiveness in curtailing the resulting foreclosure problem later in this section. As just mentioned, the structure of the mortgage instrument has changed dramatically, and we provide a brief summary of its evolution in [Section 14.4.3.1](#). Two facets of the mortgage market that have been viewed as important factors in the recent financial crisis are the efforts of the federal government to promote homeownership among low-income and minority households and the evolution of the mortgage market securitization process. We finish this section with a short discussion of their roles in the mortgage market before turning to an analysis of the causal impact of state and federal policies on the homeownership rate in the next section.

The federal government was compelled to intervene as a result of the devastating impact that the Great Depression had on the housing market. The Federal Home Loan Bank Act of 1932 and the Home Owners' Loan Corporation (HOLC) Act of 1933 were

the first responses to the problems brought on by the Great Depression. The former created the Federal Home Loan Bank Board (FHLBB) that oversaw the system of 12 Federal Home Loan Banks (FHLBanks). The FHLBanks were owned by the savings and loan associations that joined the system. One purpose of the FHLBanks was to provide liquidity to the mortgage market; members of each FHLBank could borrow from the FHLBank at low rates. To become a member, individual savings and loan associations had to purchase stock in the FHLBank. In return, they gained access to the loans and received dividends based on the amount of stock they owned.³¹ The Home Owners Loan Corporation Act established the HOLC to purchase and refinance distressed mortgages. This was in response to the high foreclosure rates that resulted from high real interest rates (due, in part, to deflation) and falling house prices during the Great Depression (Immergluck, 2009).

The bellwether policy was the National Housing Act (NHA) of 1934.³² The NHA authorized the FHA insurance program that provided a government guarantee to home mortgages. Carliner (1998) pointed out that while this did bring support to the mortgage market, the original intention of FHA was to jump-start the housing construction industry.³³ One way the FHA could do this was to make it possible for more households to purchase homes by lowering the down payment requirement; typically, it was 50–60% and FHA lowered it to 25%. To get lenders to make these loans, the FHA insured them by charging borrowers an insurance premium that went into a fund to cover default losses (Green, 2014).

After a failed attempt to establish a private secondary mortgage market that would increase the willingness of lenders to make the FHA-insured loans for this relatively novel mortgage product, the federal government created the Federal National Mortgage Association (aka Fannie Mae) in 1938 to purchase these mortgages from banks (Jaffee and Quigley, 2013). Fannie Mae could borrow money in the capital markets at reduced rates since it had the backing of the federal government. Despite Fannie Mae's recent prominence, it was not initially a big holder of FHA-insured loans (which were held primarily by insurance companies; Carliner, 1998).

Government intervention in the mortgage market was expanded in the 1944 GI Bill of Rights. This created the Veterans Administration (VA) loan guarantee program for veterans and active duty servicemen as a benefit for military service, though the VA program was also seen as a way of stimulating the housing market (Green and Wachter, 2005). VA loans were capped (as were FHA loans) though the cap did increase over time

³¹ See http://en.wikipedia.org/wiki/Federal_Home_Loan_Banks.

³² The NHA also created the Federal Savings and Loan Insurance Corporation.

³³ von Hoffman (2012) pointed out that much of the federal housing policy was strongly influenced by two competing ideological groups: the “public housers” who believed in government-funded housing and the private housing industry who believed that the government’s role should be to aid private business.

(\$4000 in 1945 and \$7500 in 1950). The big impact of the VA mortgage program was to lower the down payment needed for a loan. It substantially increased the presence of government programs in the mortgage market.

In 1940, FHA mortgages made up 13.5% of the mortgage market (in \$ amount) and this increased to 24.7% in 1945, the year the VA loans were first offered. By 1951, the two programs accounted for 44.1% of the market with VA loans accounting for 59% of the dollar value of the loans of the two programs.³⁴ Starting at the end of this decade, there was a steady decline in the share of the market attributable to the FHA and VA such that by the end of the 1990s, they accounted for less than 15% of the total dollar value of mortgages (see [Figure 14.5](#)). Part of this decline was due to gradual concentration of FHA loans in the lower income segment of the population. This arose because of the federal push in the 1960s to better serve low-income families and minorities, because of the lowering of down payment requirements, and because the FHA loan limit did not keep pace with rise in house prices to the point that it only covered households with incomes below the area median income. The decline in FHA influence was also fueled by political opposition in the 1970s to the redistributive motives of the program. Furthermore, the FHA became a part of HUD in 1965, and reorganization of HUD in 1969 and 1970 had an adverse impact on the structure and organization of FHA ([Vandell, 1995](#)). Increased competition from private mortgage insurers and the GSEs also contributed to FHA's decline.

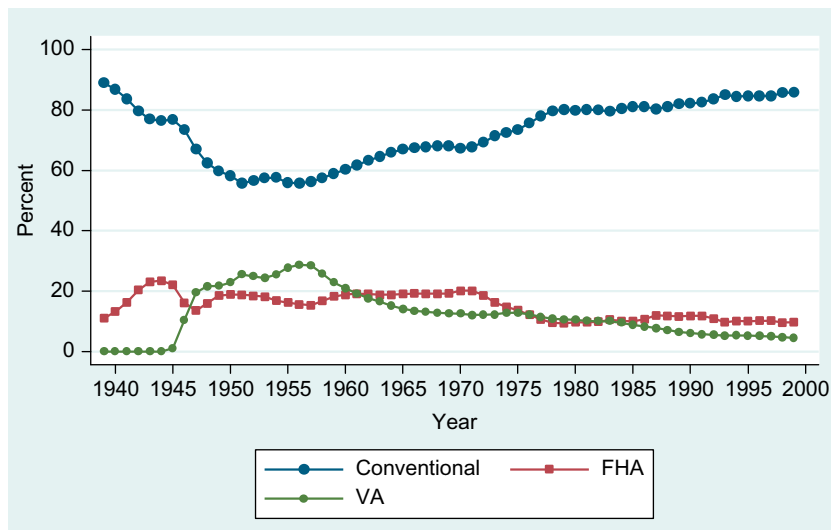


Figure 14.5 Total mortgage debt by type of mortgage written one-to-four nonfarm family homes: 1939–1999. Source: [Snowden \(2006\)](#), Series Dc929–949.

³⁴ This information comes from [Chambers et al. \(2013\)](#) who cited [Grebler et al. \(1956\)](#) as the source of the data.

Historically, the FHA has been self-financing by charging enough for its mortgage insurance to cover costs. But it faced two periods of financial troubles that also added to its demise. In 1989, an independent review by Price Waterhouse found that FHA had been underpricing its mortgage insurance for a decade, and this led to a significant decline in the FHA's Mutual Mortgage Insurance Fund. A number of government actions, including the Cranston-Gonzalez National Affordable Housing Act, led to FHA regaining actuarial soundness by 1995 (Szymanoski et al., 2012). Seller funding of down payment "gifts" from nonprofit or charitable organizations to loans with inflated house prices that allowed the sellers to recoup their money was made to households with weak credit histories who ended up defaulting at high rates. This type of loan made up 30% of FHA loans in 2005–2007 (Szymanoski et al., 2012).

Low and steady interest rates in the late 1940s through the early 1960s was a perfect scenario for commercial banks and particularly the savings and loan industry to provide mortgages. Individuals could get a better return than T-bills with little risk by depositing their money in banks, and the banks, in turn, could use these funds to purchase mortgages. The increase in the inflation rate in the second half of the 1960s led to T-bills having a higher return than what banks could pay since, under Regulation Q, the banks' interest rates were capped at a rate below the prevailing T-bill rate. This led to an exodus from banks. Furthermore, a slump in residential construction in the mid-1960s and an expected increase in housing demand spurred by baby boomers led to the HUD Act of 1968. This act made Fannie Mae a private company with stock listed on the New York Stock Exchange (NYSE). Fannie Mae could now purchase and sell conventional nongovernment-insured mortgages. An additional motivation for making Fannie Mae a private corporation was that the government was able to remove Fannie Mae's debt from its balance sheet. At the same time, a new agency, the Government National Mortgage Association (Ginnie Mae), was created to continue purchasing and securitizing loans insured by the FHA and the VA.

Fannie Mae still retained ties to the federal government due to the federal government's implicit guarantee of loans that it purchased. It also has other privileges such as exemption from state and local taxes. In return for having these benefits, Fannie Mae was obliged to assist mortgage funding for low- and moderate- income families and underserved parts of urban areas.

Two years later, at the behest of the thrifts who wanted similar status as the commercial banks, the Emergency Home Finance Act (EHFA) of 1970 was passed. EHFA created the Federal Home Loan Mortgage Corporation (aka Freddie Mac) under the control of the FHLBB. Originally, Freddie Mac was owned by the 12 FHLBanks and by S&Ls that were members of the FHLB system. It was privatized in 1989 with stock listed on the NYSE but with the same special privileges as Fannie Mae. Together, Fannie Mae and Freddie Mac are referred to as GSEs.

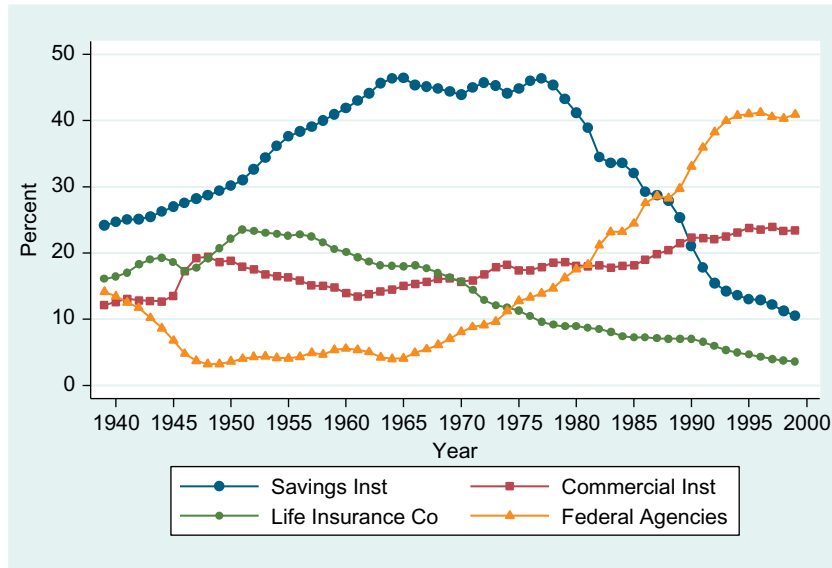


Figure 14.6 Percent of total mortgage debt by holder: 1939–1999. Source: Snowden (2006), Series Dc929–949.

Figure 14.6 plots the market shares of total outstanding mortgage debt held or securitized by savings institutions, commercial banks, life insurance companies, and federal and federally related agencies. The latter include FHA, VA, and the GSEs. One can see that the federal and federally related agencies had the smallest holdings of these four groups until 1975. Over the next 20 years, there was a dramatic rise in the market share held or securitized by federal agencies from around 10% to 40%. As of 1999, they held 41% of total outstanding mortgage debt and commercial banks held the next largest share of 23%. The combined GSE share of outstanding mortgage debt was around 25% in 1990 and grew to more than 46% in 2003. Over the next 3 years, its share declined to less than 39% due to the increased activity by private entities (see Figure 14.7).

14.4.3.1 The evolving structure of mortgages

The structure of mortgages has changed dramatically since HOLC was established in 1933 and the FHA was first formed in 1934. Table 14.2 provides information on the structure of mortgages for one-unit, owner-occupied properties starting in 1920. Prior to FHA, mortgages were typically short term (5–10 years), with no amortization and with a balloon payment at the end. The typical LTV was 50–60%. These loans were usually refinanced on a regular basis so as to avoid the balloon payment. But the Great Depression resulted in falling house prices (for example, see figure 3 in Rose, 2011) and lenders were unwilling to refinance loans, which led to a dramatic rise in foreclosures and a huge contraction in the mortgage market.

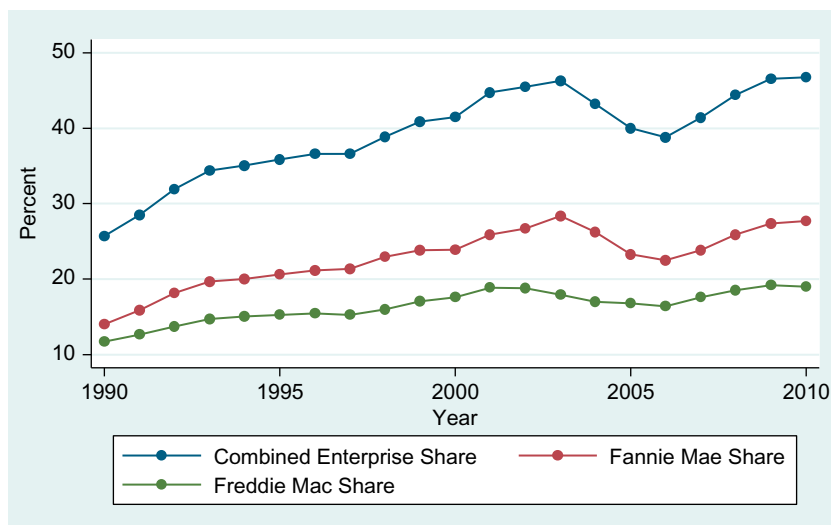


Figure 14.7 Fannie Mae and Freddie Mac's share of outstanding mortgage debt: 1990–2010.
Source: FHFA.

Table 14.2 Loan terms on one-unit, owner-occupied properties

	1920	1950	1960	1970	1990	2000
Median LTV (%)	50–60	75	79	84	85	91
VA		91	91	95	100	100+
FHA		79	83	93	100	99
Conventional		66	68	77	77	94
Percent with LTV ≥ 100		12	9	14	29	29.5
VA		32	20	34	40	64
FHA		1	3	12	32	42
Conventional		9	6	10	27	12
Median loan term (years)	5–11	13	20	25	30	30+
VA		20	25	29	30	30+
FHA		20	24	29	30	30+
Conventional		11	15	21	25	30+
Median interest rate	6–7	5	5.1	6	9.5–9.9	7.1
VA		4	4.5	5.4	9.1–9.4	7.9
FHA		4.5	4.6	5.8	9.5–9.9	8.0
Conventional		5	5.6	6	9.5–9.9	7.2

Notes: Data for 1920–1970 are from Fetter (2013a). Data for 1990 and 2000 are from the US Census Bureau's Residential Finance Survey and represent the stock of first mortgages. Loan-to-value (LTV) is defined as the amount of the first mortgage loan as a percent of purchase price, for properties acquired by purchase with first mortgage made or assumed at time of purchase. Conventional loans are mortgages without government insurance.

One of FHA's largest influences was the popularization of a standardized mortgage instrument—a 20-year, fully amortizing loan with a 20% down payment requirement. But as [Carliner \(1998\)](#) noted, the initial features of the FHA such as strict appraisals, new standards of construction and design, and escrow of tax and insurance payments were implemented to reduce risk rather than increase the homeownership rate.

As [Rose and Snowden \(2013\)](#) documented, though, the FHA did not create the fully amortizing loan. It was first introduced to the United States through the buildings and loan (B&L) industry in general and specifically through the B&L institutions in Dayton, Ohio, in the 1870s and 1880s. The gradual adoption of the amortizing loan was due to innovations that made it beneficial from a cost–benefit standpoint. But it was not until the 1930s when the more common loan structure known as the share accumulation contract resulted in high failure rates that the amortizing loan became significantly more popular in the B&L industry. Rose and Snowden pointed out that this transformation was not directly due to FHA and the HOLC since the B&Ls did not make much use of the FHA insurance program. It is likely that FHA provided indirect pressure in the form of competition for loans with the B&Ls.

As of 1950, VA and FHA loans tended to be longer term (e.g., 20 or 25 years) and to have lower down payments and interest rates (VA loans often had no down payment) than conventional mortgages. The median LTV for VA mortgages was 91; for FHA, it was around 80; and for conventional mortgages, it was 66. The median loan term was 20 years for VA and FHA mortgages and 11 years for conventional mortgages. There was a slow upward trajectory in LTV over the next 40 years, particularly in FHA mortgages, so that by 1990, the median LTV was 100 for both VA and FHA mortgages and was 77 for conventional mortgages. By 1990, the median loan term increased to 30 years for VA and FHA mortgages and to 25 for conventional mortgages.

Prior to 1980, a very large percentage of mortgages were fixed-rate mortgages. The high inflation rates of the early 1980s led to major problems for the savings and loan industry that was saddled with interest rate ceilings. In order to provide relief, the federal government deregulated the S&Ls by abolishing interest rate ceilings on first-lien residential mortgages (Depository Institutions Deregulation and Monetary Control Act (DIDMCA) of 1980) and allowing for adjustable rate mortgages (ARMs) (Garn–St. Germain Depository Institutions Act of 1982). ARMs typically have a lower initial “teaser rate” that adjusts to a higher rate at a prespecified point in the life of the loan. One version that dates back to this period is called the payment-option ARM with one of multiple options that allows for a monthly payment that is less than the interest due on the loan. This leads to an increase in the principal of the mortgage. This became a popular product in parts of the country. For example, one-third of all loans in California in 1996 were payment-option ARMs ([Foote et al., 2012](#)). The obvious advantage to the borrower was the lower initial payments that made it more affordable in the short run. The impact of this deregulation on the structure of mortgages was focused on the

conventional loan market where the median LTV increased from 77 in 1990 to 94 in 2000 (Table 14.2). While the intention was aimed at revitalizing the savings and loan industry, many see these deregulations as setting the stage for the subprime mortgage crisis in the 2000s (Office of Policy Development and Research of the US Department of Housing and Urban Development, 2010).

The evolution of the mortgage market after the deregulation in the early 1980s was a slow and gradual process. To a large extent, it is characterized by the evolution of the subprime market that did not experience significant growth until the mid-1990s when there was a surge in refinance loans (subprime purchase loans also increased but at a slower rate). These loans were typically used to cash out on home equity rather than to lower interest rates. They tended to be 30-year fixed-rate mortgages with higher rates, fees, and prepayment penalties than prime mortgages (Immergluck, 2009). In response to the increase in the origination of subprime loans, Congress passed the Home Ownership and Equity Protection Act (HOEPA) in 1994. HOEPA imposed significant restrictions on the lending terms and practices of “high-cost” loans though it only covered approximately 5% of subprime loans (Bostic et al., 2008). Surprisingly, Ho and Pennington-Cross (2007) and Bostic et al. (2008) found evidence that such restrictions can actually increase subprime lending since they reduce borrowers’ fears of predation and hence increase the likelihood of subprime originations.

Starting in 2001, there was a large increase in subprime purchase loans; the number almost doubled from 1.1 million in 2003 to 1.9 million in 2005. Furthermore, Alt-A loan originations more than tripled over this period such that, in dollar terms, nonprime loans made up 32% of all mortgage originations in 2005 (Mayer et al., 2009).³⁵ These subprime loans tended to be ARMs with low down payments, much higher debt-to-income ratios, and features such as no- or low-documentation, interest-only, and negative amortization. Another new type of mortgage was known as the piggyback loan that was used to avoid paying private mortgage insurance. Buyers would take out a first mortgage with an LTV of 80 and then finance the down payment with a second loan. So, effectively, the LTV was 100.

Mayer et al. (2009) analyzed loan performance data for the 2003–2007 period. They found that 75% of the subprime mortgages in their sample were ARMs with lower “teaser” rates for the first 2 years. Alt-A loans consisted of a substantially higher percentage of fixed-rate mortgages. The share of subprime piggyback loans increased from 7% in 2003 to 28% in 2006, whereas this share for Alt-A loans increased from 12% in 2003 to 42% in 2006. About a third of subprime loans were no- or low-documentation loans, whereas more than two-thirds of Alt-A loans had this property.

³⁵ Subprime mortgages are typically made to individuals with poor credit ratings (FICO score below 640) though the working definition of subprime mortgages has changed over time. Alt-A mortgages are considered to be riskier than conforming mortgages that are purchased by the GSEs. A typical reason is that there is no borrower income documentation.

Mayer et al. (2009) found little evidence that the significant rise in delinquencies and defaults starting around the beginning of 2007 was linked to the proliferation of new and novel mortgages such as ARMs with initial teaser rates, mortgages with prepayment penalties, negative amortization, and interest-only mortgages. Rather, the authors attribute the higher defaults to the deterioration in lending standards. First, there was a large increase in subprime mortgages that defaulted soon after origination. Between 2000 and 2004, 1.5% of subprime mortgages defaulted within a year of origination, whereas 12% of 2007 originations did so. That is, underwriting standards fell so dramatically that borrowers were essentially unable to make payments right from the start. Second, the higher LTV ratios in conjunction with falling house prices meant that borrowers were quickly underwater and this increased their motivation to default. These risky borrowers were unlikely to have been approved for more standard fixed-rate, prime mortgages with larger down payments. As we discuss below, the rise in defaults and ultimately foreclosures likely led to the significant drop in the homeownership rate starting in 2009.

14.4.3.2 Low-income and minority homeownership policy

The homeownership rate for low-income households has always been significantly lower than that for higher-income households (see [Figure 14.4](#)). The federal government has tailored some of its housing policies toward reducing this gap by promoting homeownership among low-income households. The rationale for this can be explained by the general belief in homeownership as the American Dream. It was also felt that promoting homeownership for low-income households that meets a minimum housing standard can improve on behavioral, social, economic, and neighborhood conditions that are particularly bad for this group ([Shlay, 2006](#)).

Starting in 1938, the FHA began to insure 25-year loans with a maximum LTV of 90% to low-income households as a way of expanding homeownership opportunities to this segment of society. The cap on these loans was set at \$5400, whereas the initial cap for FHA-insured loans was set at \$16,000, which was well above the average house value at the time. The number of FHA-insured loans that went to low-income households was minimal until the 1960s. This can be explained, in part, by the FHA underwriting standards that limited loans in minority areas. These racially biased standards were overturned by President Kennedy in 1962 ([Carliner, 1998](#)).

As of the early 1960s, the federal government had done little to promote homeownership among low-income households. This changed with Section 235 of the HUD Act of 1968 that subsidized the construction of new housing to be sold to low-income households and also subsidized loans to low-income households for existing properties. There was a negligible down payment and annual payments were limited to a maximum of 20% of income or 1% annual interest.

With a large goal of building 26 million units over 10 years, construction was expedited and FHA oversight was corrupt as substandard units were built and sold at inflated

prices. This resulted in high foreclosure rates (Olsen and Ludwig, 2013). Despite the accelerated process, Section 235 only financed around 400,000 homes for low- and moderate-income families by the beginning of 1973. One reason for the low uptake was that the main source of information about the program was the real estate industry and not the local offices of the FHA (Olsen, 2007a). Furthermore, Section 235 was criticized for not serving the truly needy. Section 235 was shutdown in 1973 due to the scandals surrounding its implementation. It was reinstated in 1976 with a higher interest rate and larger down payment and resulted in about 125,000 loans before it was permanently ended in 1987.

A more successful though less well-known federal program that subsidized low-income homeownership was a provision of the Farmers Home Administration Section 502 program that helped provide mortgages in rural areas. This provision was also enacted in 1968 and provided subsidies to rural households with incomes below 80% of local area median family income. More than 1 million loans were originated under this program between 1969 and 1993 (Carliner, 1998). This was almost twice as many as under Section 235.

A federal block grant program that focuses on providing funding for low- and moderate-income housing is the HOME Investment Partnership Program. It was authorized as Title II of the Cranston-Gonzalez National Affordable Housing Act of 1990 and allocates about \$2 billion a year in federal funds to state and local governments. In 2002, state and local governments devoted about 48% of their HOME budget to homeownership assistance in the form of subsidized loans for home rehabilitation and homebuyer assistance. Through 2002, HOME subsidized more than a quarter-million homebuyers (Turnham et al., 2004).

Racial discrimination and redlining (purposely limiting financing in areas based on their racial characteristics) has had a long history in the US housing market. If anything, the FHA promoted discriminatory lending practices through the use of an underwriting standard of neighborhood (racial) homogeneity that was typically enforced through restrictive covenants. Despite being outlawed by the Fair Housing Act of 1968 (Section VIII of the Civil Rights Act of 1968), evidence emerged in the early 1970s that banks were still redlining and that minorities were being denied credit (e.g., Munnell et al., 1996). Access to credit was limited in low-income and high-minority areas as banks were taking deposits in branches in these areas and then originating loans elsewhere. All this motivated the Home Mortgage Disclosure Act (HMDA) that was enacted in 1975 and the CRA that was passed in 1977 under President Carter. The CRA stipulated that commercial banks and thrifts with branches in moderate- and low-income areas meet the needs of borrowers in these areas. These institutions are periodically (every 2 or 5 years) audited to see if they are meeting the CRA goals. The outcome of the evaluation is a rating of “outstanding,” “satisfactory,” “needs to improve,” or “substantial noncompliance.” The motivation for banks to show reasonable

performance is that this rating is considered when banks apply for mergers, charters, deposit insurance, or office relocation.

In the late 1980s and early 1990s, Fannie Mae and Freddie Mac were also seen to be encouraging lending in higher-income areas while making it more difficult for loans to be originated in low-income neighborhoods through its underwriting practices (Immergluck, 2009). This led to the Federal Housing Enterprises Financial Safety and Soundness Act of 1992 (the 1992 GSE Act) that mandated the GSEs to purchase a certain percentage of loans originated in low- to moderate-income tracts or to low- to moderate-income borrowers. The GSE Act also set up a new entity within HUD, the Office of Federal Housing Enterprise Oversight (OFHEO), to provide regulatory oversight of the GSEs.

How effective was the CRA and the 1992 GSE Act in promoting minority and low-income lending? Recall that the homeownership rate for blacks increased by around 7 percentage points between 1995 and 2004 but then declined by 6 percentage points to a level that was actually lower than the rate in 1980. The homeownership rate for households with family income below the median decreased by 2 percentage points between 1990 and 1994 before increasing by 4.5 percentage points by 2005 but then falling back by 2013 to the same level as in 1990. Can any of the increases in the minority and low-income homeownership rates be attributed to the CRA and GSE affordable housing goals?

Initially, the CRA was fairly ineffective (very few banks received a grade below “satisfactory”). In light of this, the 1989 Financial Institutions Reform, Recovery, and Enforcement Act required the public disclosure of the CRA exam and the CRA evaluation process was changed in 1995 to make it more performance-oriented and less subjective. Bhutta (2011) did find that CRA resulted in a 7% increase in loan volume for large banks in large MSAs between 1997 and 2002. The rationale is that large banks are more likely to merge or to open new branches and the CRA rating is important for doing so. Bhutta also found that there was an increase in lending by bank subsidiary mortgage companies and independent mortgage companies not covered by CRA. His explanation is that the CRA lending resulted in information externalities in historically thin markets (see Section 14.4.1) that allowed these banks to reduce their costs and hence increase lending.

One characteristic of CRA that limited its impact is that it only applies to depository institutions. The dramatic decline in the S&Ls after the S&L crisis in the late 1980s and the large increase in nondepository mortgage companies that took their place meant that CRA applied to a percentage of the mortgage market that got smaller over time. The share of originations made by CRA-covered institutions was around 80% in 1977 and approximately 65% in 1990 and 45% in 1996 (Immergluck, 2009). This was followed by a large increase in subprime loans that were mostly originated by nondepository institutions. In 2005 and 2006, only 6% of subprime mortgage originations qualified under CRA (Bhutta and Canner, 2009).

After the GSE Act, the GSEs did show improvement in lending to low-income and minority census tracts (Immergluck, 2009). But as will be discussed later, there is little evidence to support the increased supply of mortgages in census tracts that qualify under the GSE affordable housing goals. This is explained, in part, by the high levels of GSE crowd out (Gabriel and Rosenthal, 2010).

14.4.3.3 Securitization

In 1970, Ginnie Mae issued the first mortgage-backed security (MBS) and Freddie Mac was not far behind in 1971. The owners received the payment of principal and interest from the mortgages that made up the MBSs. These MBSs had the backing of Ginnie Mae and Freddie Mac and hence government guarantees which the owners paid for with a small premium. The creation of the MBSs allowed Ginnie Mae and Freddie Mac and ultimately Fannie Mae to pass on the interest rate risk though they kept the credit risk by guaranteeing the MBSs.

Private firms, typically nonbank mortgage companies, first entered the secondary market in 1977. Their MBSs tended to be made up of “nonconforming” mortgages such as jumbo mortgages with values above the conforming loan limit that were not purchased by Fannie Mae or Freddie Mac. These were referred to as private-label mortgage-backed securities (PLMBSs). The PLMBS market really took off in the beginning of the 2000s through securities made up of subprime and Alt-A mortgages. Given that PLMBSs did not have the same implicit government backing of MBSs sold by the GSEs, they were structured to handle credit risk through the generation of varying risk levels or tranches with associated bonds with different credit ratings. These PLMBSs were generally not regulated since they were issued by nonbank mortgage companies (Ellen et al., 2011).

The evolution of the MBS market was initially quite slow. But momentum was provided by a number of federal policies. These included the aforementioned DIDMCA of 1980 and the Garn–St. Germain Depository Institutions Act of 1982. Along with lifting the caps on interest rates, DIDMCA allowed national banks to only be subject to the interest rate rules of their home state and, in essence, apply rules from low-regulation states. This gave them an advantage over local banks that were subject to local (and potentially stricter) regulations. The Garn–St. Germain Depository Institutions Act overrode state laws in allowing for ARMs and other alternative mortgage types. It also allowed formerly state-regulated mortgage companies to choose to be overseen by the federal S&L regulator. Finally, in the wake of the S&L crisis in the late 1980s, the 1989 Financial Institutions Reform, Recovery, and Enforcement Act dramatically changed the savings and loan industry and its federal regulation. All these new rules fostered the growth of national banks and mortgage companies (at the expense of S&Ls) that tended to rely on securitization as a source of funding (as opposed to deposits) (Immergluck, 2009).

The total value of MBSs issued in 1985 was \$100 billion; in 1995, it was a little more than \$250 billion with almost all issued by the GSEs. This increase in securitization by the

GSEs helps to explain their increasing share of the mortgage market in the 1990s. The MBS market then started to grow rapidly in the 2000s; it peaked in 2003 with more than \$2.5 trillion in issuances, the vast majority by the GSEs, but in 2005, there was a little more than \$2 trillion in issuances and less than half was by the GSEs. Of course, by 2008, PLMBS issuances were practically zero.³⁶

The GSEs started purchasing PLMBS in the early 2000s. They held around \$100 billion (\$2009) of PLMBS in 2002 and increased their holdings to more than \$350 billion (\$2009) of PLMBS in 2005 (CBO, 2010). The GSEs could meet their goals by purchasing PLMBSs that included mortgages that met the GSE affordable housing goals. In fact, mortgages can satisfy more than one goal and PLMBSs with such mortgages were desirable.

14.4.4 The impact of federal intervention in the mortgage market on homeownership rates

We first focus on two episodes of significant changes in the homeownership rate, 1940–1960 and 1994–2009, and to what extent government housing policies played a role in the change in the homeownership rate during these periods. Next, we discuss government policies to promote low-income homeownership and consider to what extent they played a role in the recent financial and housing crisis. This is followed by an analysis of research that estimates the pass-through of the GSE subsidy to lower mortgage interest rates and the subsequent impact on homeownership. Finally we evaluate the impact of federal policies and state laws on mortgage default and foreclosure rates.

14.4.4.1 Explaining the postwar surge in homeownership

The large increase in the homeownership rate between 1940 and 1960 has been linked to a number of underlying factors such as changes in demographics, the rise in real income, and changes in the mortgage market. But Fetter (2013b) argued that the very large increase in the subperiod of 1940–1945 that amounted to half the total increase over the 1940–1960 period was due primarily to a government policy that was not intended to promote homeownership, namely, the imposition of rent control. The motivation for rent control was the increase in rents that resulted from the huge inflows of workers in some areas of the country that were involved in the ramping up of military production for World War II (WWII). The federal government felt it was necessary to cap rents; otherwise, the large increases could lead to rises in wages and other prices that could dampen the war effort. This regulation was pervasive since, as Fetter says, it was eventually imposed on 80% of the rental stock in the United States. This cap on rents along with increases in house prices (that were not regulated) provided incentives for landlords to convert rental units to owner-occupied housing. Furthermore, one source of the increase

³⁶ The source of this information is the Securities Industry and Financial Markets Association's statistics on structured finance.

in homeownership, the building of new homes, was not a factor during this time since new construction was limited.

Rent control capped prices based on the existing rents prior to the start of the war buildup. Since the buildup started at different times in different places, there was variation in the date that determined the rent limit (the base date). Rent control was imposed 60 days after an area was designated as a “defense rental area.” After an initial period in which surveys of rents were used to determine the areas where increasing rents threatened the buildup effort, there was a blanket imposition of rent control on most of the rest of the country in October 1942.

Data used are rent indexes for 51 cities starting in March 1940 that come from the National Industrial Conference Board along with housing surveys from the Census Bureau and BLS for 1944–1947. Fetter regressed the change in the homeownership rate between April 1940 and the date of the BLS housing survey (sites were surveyed once during the 1944–1947 period) on the severity of the rent control and the percent decline in rents between the maximum precontrol value and the initial value at the base date. Identification arises from the variation in the severity of the rent control that comes from the variation in the base date, the subsequent rise in rents prior to the imposition of rent control, and the ultimate fall in rents after control. An important control is the maximum precontrol rent appreciation (from March 1940 to the precontrol maximum rent) so that the source of identification is between cities with similar precontrol rent appreciation but differing levels of rent control severity based on the actual decline in rents. Through the addition of other controls and numerous robustness checks, Fetter showed that the rent control severity variable is uncorrelated with other unobservables that could affect the homeownership rate. The results indicate that rent control can explain 65% of the increase in homeownership over the early 1940s.

Three recent papers look at the impact of the VA Mortgage program on the large increase in the homeownership rate between 1945 and 1960 (Vigdor, 2006; Chambers et al., 2013; Fetter, 2013a).

Vigdor (2006) pointed out that the main benefit that the VA Mortgage program provided for eligible veterans was to lower their down payment requirement (possibly to zero). This can be seen as a relaxation of credit constraints, which makes it easier to get a mortgage and hence increase homeownership. Using Integrated Public Use Microdata Series (IPUMS) data for 1940, 1970, 1980, 1990, and 2000, Vigdor showed that being an eligible veteran increases the likelihood of being a homeowner in 1970 by 0.07—a semielasticity of 12.2%. He concludes that since 38% of households were eligible for VA Program benefits, this impact accounts for about 20% of the increase in homeownership between 1940 and 1970.

To see if this easier access to credit affected house prices, Vigdor ran a regression of owner’s house valuation (single-family detached) on the median MSA rent, the share of veterans in the MSA, and median rent interacted with the share of veterans using the

IPUMS data from the 1970 decennial census. The result is that a one standard deviation increase in the percentage of veterans increases the value-rent multiplier by 40. Vigdor pointed out that the national increase in the value-rent multiplier between 1940 and 1970 was approximately 60 so this is a large impact.

Given that this result is based on a cross-sectional regression, it is not clear that the variation in the share of veterans interacted with median rent is truly exogenous; it is likely that there are unobservables that are correlated with house prices, median rents, and the interaction of median rents and share of veterans that would bias the results.

Fetter's (2013a) approach is similar to Vigdor's, but he used a RD framework as a means for providing exogenous variation in veteran's status to estimate the causal impact of VA Mortgage program eligibility on homeownership. The VA program is eligible to veterans and Fetter used the wind down in WWII in 1945 and in the Korean War in 1953 as an exogenous change in military service at the initial draft eligibility age to obtain causal estimates. One issue is there is not an official end to the draft for either war (so this is really a fuzzy RD), and so Fetter had to estimate the discontinuity using a structural break procedure: date of birth (DOB) 1 January 1928, for WWII and DOB 1 October 1933, for the Korean War. Similar to Vigdor, Fetter used IPUMS data from 1960, 1970, and 1980. The IV estimates of the impact on homeownership in 1960 are 0.129 and 0.177 for WWII and Korean War veterans, respectively. The corresponding semielasticities are large: 30% and 80%.³⁷

What is interesting is that the impacts on homeownership in 1970 and 1980 are not significantly different from zero. Given that the age at the break point in 1960 is 32 for WWII veterans and 26 for Korean War veterans, this supports the result that VA eligibility caused a shift in homeownership to an earlier age. The reasoning for the age shift is that VA mortgages relaxed the down payment requirements and younger households are more income-/wealth-constrained and hence are more responsive to the lowering of the down payment requirement.³⁸ Note that the nonsignificance in 1970 is in contrast to Vigdor who found that eligible veterans were significantly more likely to own a home. This is evidence that the variation in the share of veterans used in Vigdor's study is probably not exogenous.

To determine the overall impact on homeownership, Fetter extrapolated the estimated impacts at ages 26 and 32 obtained from the RD analysis to the full age distribution. He found that the VA program accounts for 7.4% of the increase in the homeownership rate between 1940 and 1960. Fetter noted that one of the broad changes

³⁷ Fetter ruled out other factors related to military service that might affect homeownership such as differences in education, income, and military service itself (using WWI veterans).

³⁸ Using data from the PSID and NLSY from 1984 to 1990, Goodman and Nichols (1997) also provided evidence that the lower qualifying standards for FHA-insured mortgages just shifted the age at which households became homeowners rather than increasing the number of homeowners.

in the mortgage market over the midcentury period was the reduction in the down payment from around 40–50% to 20% or less (see [Table 14.2](#)). Using the results from the VA analysis, Fetter estimated that if all homeowners in 1960 had been required to make a 50% down payment, the homeownership rate would have been 11 percentage points lower or 40% of the increase in homeownership between 1940 and 1960.

[Chambers et al. \(2013\)](#) noted that the Case–Shiller house price index increased by 41.4% between 1935 and 1960 (this was actually less than the increase in real income over the same period). Their goal is to explain the comovement between the home ownership rate and house prices in the postwar housing boom. The authors develop a general equilibrium three-sector model with housing, mortgage markets, and consumptions goods that includes both ownership and rental options. They calibrate the model using data from 1935 to 1940.

Chambers et al. showed that demographics, income risk, and government housing finance are key determinants of the increase in the homeownership rate between 1940 and 1960. In particular, the latter is due to the change from a balloon payment contract to a longer-term mortgage with a lower down payment. Hence, all three papers provide evidence that federal programs contributed to the large increase in homeownership during the 1945–1960 period.

On the other hand, Chambers et al. found that it was the rise in construction costs and not government housing policy that led to the increase in house prices. This is in contrast to Vigdor who found that the VA Mortgage program had a large impact on house prices. This brings into question the exogenous variation in veteran's status that Vigdor used to identify the price effect.

14.4.4.2 The recent rise in homeownership

After several decades of stability, the US homeownership rate increased from 64% in 1994 to 69% in 2005. To explain this increase, [Chambers et al. \(2009; CGS hereafter\)](#) first showed that a large part of the overall increase in the homeownership rate during this period was due to the increase from 37% to 43% for households of age 35 years and under. There were also significant increases in the homeownership rate for the second and third income quintiles. This motivated the authors to look at factors that reduce the financial burden of becoming a homeowner as a potential source of these increasing participation rates.

To do so, CGS developed a general equilibrium overlapping generations model for housing and mortgage markets that includes both ownership and rental options. They calibrated the model using data from 1994. CGS used the model to decompose the gain in the homeownership rate over the next decade into components that are due to demographic changes and innovations in the mortgage market. First, they showed that reductions in the transaction costs of buying property and decreases in down payment requirements had little impact on the homeownership rate. What appears to have had

a significant impact was the introduction of piggyback loans where the second loan covers the down payment (in conjunction with the standard fixed-rate mortgage with a 20% down payment). In the model that allows for the combined effects of demographics and mortgage innovation in accounting for the observed change in homeownership, the introduction of the piggyback loan accounts for between 56% and 70% of the increase, whereas demographics account for the remainder. The new product is particularly attractive to the 20–34 age group, which explains their large increase in homeownership over the 1994–2005 period.

Gabriel and Rosenthal (2015) investigate the factors that determine the rise and fall of the homeownership rate in the 2000s. They use household-level IPUMS data from the 2000 decennial census and the 2005 and 2009 American Community Surveys. The authors run a reduced-form regression of tenure status on a host of demographic, labor/employment, and geographic controls and three measures of local housing conditions. The latter include the median owner-assessed value of owner-occupied homes for the public use micro area of residence, the 1-year-ahead forecast of house prices in the MSA, and a measure of house price volatility in the MSA. Separate regressions are run for each year and age between 21 and 89 using the full sample and for those households that moved within the previous 12 months.

The authors use a shift-share methodology to decompose differences in homeownership rates across years into changes in sociodemographic factors and market conditions. They show that the latter were the key factors in the increase in homeownership rates between 2000 and 2005 and in the decline between 2005 and 2009. They note that there is indirect evidence that changes in underwriting standards played an important role in these changes. That is, the loosening of standards helped push up homeownership rates in the beginning of the decade and the subsequent tightening of mortgage supply helped push down the rate in the latter half of the decade.

In summary, both CGS and Gabriel and Rosenthal provide evidence that easier access to mortgage credit was an important contribution to the rise in the homeownership rate between 1994 and 2004.

14.4.4.3 The impact of the affordable housing goals

Continued evidence of redlining in the 1960s and 1970s led the federal government to pass laws aimed at increasing lending to low-income households. The most important of these laws were the CRA in 1977 and the GSE Act of 1992. These laws mandated qualifying banks and the GSEs to meet minimum purchase requirements of mortgages held by low-income and minority households. These are referred to as the *affordable housing goals*.

For qualifying banks, the most common way of meeting the CRA goals was by originating or buying residential mortgages for properties in low- to moderate-income census tracts, those where median family income is less than 80% of area median income in their assessment areas (usually counties in which they have deposit-taking offices/branches).

Loans to low-to-moderate-income borrowers also qualify for meeting the CRA goals. Low-to-moderate-income status is determined using the decennial census. Independent mortgage banks and credit unions are not covered under CRA. Plus, more than half of loans made or purchased by CRA-covered institutions were made outside of their assessment areas, further limiting the number of loans covered under CRA (Avery and Brevoort, 2011).

HUD sets the affordable housing goals for the GSEs related to (1) low- and moderate-income families, (2) purchasers of properties located in historically underserved areas (underserved area goal), and (3) low-income families living in low-income areas and very low-income families (the “special affordable” goal). The low- and moderate-income goal defines a low- or moderate-income household as one whose income is less than or equal to the area median household income. For metropolitan areas, “underserved areas” are defined as census tracts with either (1) at least 30% minority population and with a median family income at or below 120% of the area median family income or (2) a median family income at or below 90% of the area median family income. The “special affordable goal” defines a very-low-income household as a household whose income is less than or equal to 60% of the area median income. For a list of the targets for these three affordable housing mandates, see An et al. (2007).

In this subsection, we look at the effectiveness of the CRA and GSE affordable housing goals in increasing the homeownership rate of low-income households. We also consider indirect evidence about how the affordable housing goals affected loan volume since this is a necessary condition for there to be an effect on the overall homeownership rate. The issue of loan quality is also discussed since this has received so much attention in relation to the financial crisis and because it is tied in so closely with credit supply.

Some studies blame the CRA and the GSEs for their role in perpetrating the financial crisis by motivating originators to lower their standards and extend credit to risky borrowers to meet the affordable housing goals.³⁹ To put this issue in context, as reported by Bhutta and Canner (2009) using HMDA data for 2005 and 2006, only 6% of subprime mortgage originations qualified under CRA and the performance of CRA-related subprime loans was similar to other subprime loans. For the most part, subprime mortgages were not conforming loans so they were not eligible to be directly purchased by the GSEs whether or not they could be used to meet their affordable housing goals. So the task is to show how CRA and the GSEs could have had such a large impact on the financial crisis given that they were involved in such a small percentage of subprime originations. The GSEs did purchase PLMBS tranches that included subprime mortgages that could be used

³⁹ Studies that claim to provide evidence that CRA and the GSEs helped cause the financial crisis include Liebowitz (2009), Wallison (2009), and Nichols et al. (2011). But Liebowitz (2009) and Wallison (2009) only provided circumstantial evidence and the analysis in Nichols et al. (2011) clearly suffers from endogeneity problems.

to meet their affordable housing goals. We will also look at this avenue as a potential way that the GSEs could have contributed to the financial crisis.

Credible investigations into the impact of the affordable housing goals on the homeownership rate and their role in the financial crisis require controlling for the numerous sources of omitted variable bias that can contaminate the results. Hence, the evidence on the impact of the affordable housing goals on homeownership, loan supply, and loan quality to which we give the most weight is based on some application of the quasi-experimental methodology.⁴⁰

A common approach to coming up with plausible causal estimates of the affordable housing goals on the homeownership rate, loan volume, and loan performance is to use their cutoffs in terms of census tract median household income relative to area median income or individual household income relative to area median income as the basis for a RD analysis or as a source of exogenous variation to construct valid instruments for GSE or CRA activity. In the rest of this subsection, we discuss such estimates of the impact of the affordable housing goals on the homeownership rate, loan volume, and loan quality.

Papers that estimate the impact of the affordable housing goals on the homeownership rate include [Bostic and Gabriel \(2006\)](#), [An et al. \(2007; henceforth ABDG\)](#), and [Gabriel and Rosenthal \(2009\)](#). All three papers use 1990 and 2000 decennial census data at the census tract level and the latter two use tract-level HMDA data (ABDG from 1995 to 2000 and Gabriel and Rosenthal from 2000). All three use an informal RD approach to compare CRA and/or GSE activity in census tracts just on either side of the affordable housing goal cutoffs (e.g., census tracts with median family income that is 80% and 90% of area median income for CRA and the GSEs, respectively). For the CRA target, Gabriel and Rosenthal found that there is a positive and significant impact on the supply of nonconforming loans and limited evidence of a positive impact on the homeownership rate. They claimed that this is evidence of the effectiveness of CRA in increasing mortgage supply in targeted areas. Otherwise, the results show little evidence that GSE eligibility had any effect on changes in homeownership rates.

ABDG looked more directly at the GSE effect by using the percent of mortgages in a census tract that were purchased by GSEs and the change in this variable as explanatory variables. They instrumented for these two variables using the GSE target indicators and the total number of conforming loans in 1995. They found that GSE intensity has a significantly positive impact on the change in the homeownership rate and the percent

⁴⁰ There are numerous studies that do not use such techniques to identify the causal effects of the Affordable Housing Goals on relevant outcomes and hence are likely to suffer from omitted variable bias. These include [Quercia et al. \(2003\)](#), [Ambrose and Thibodeau \(2004\)](#), [Demyanyk and Van Hemert \(2011\)](#), and [Reid and Laderman \(2011\)](#).

change in GSE intensity had a significantly negative impact on vacancy rates and a significantly positive impact on house prices. This is evidence that GSE intensity is related to neighborhood improvements.

Gabriel and Rosenthal (2010), Avery and Brevoort (2011), Bhutta (2011, 2012), Ghent et al. (2013), Moulton (2014), and Bolotnyy (2014) are the best studies that attempt to show a causal impact of the affordable housing goals on loan volume (subprime or prime). The central theme of these papers is to use the affordable housing goals in a RD framework to estimate their causal impact on the quantity and quality of loans. The main takeaway from these studies is that, for the period from the mid-1990s to 2007, there is little evidence that the affordable housing goals had a significant impact on any of these factors. Agarwal et al. (2012b) estimated that CRA led to a decline in loan quality using plausibly exogenous variation in banks' incentives to meet CRA goals around regulatory exams. But Reid et al. (2013) and Foote et al. (2013) criticized the exogeneity assumption used by these authors to obtain their estimates.

One reason for the lack of an impact of the affordable housing goals on the mortgage market is crowd out. That is, the activities of the GSEs just displace mortgage supply that would have occurred in their absence. Gabriel and Rosenthal (2010) showed that crowd out in the home purchase market (most relevant for the homeownership rate) is positively related to the level of market activity; it was most prevalent during the 2003–2006 period when it reached the 50% level. Little crowd out existed in periods prior to and subsequent to this market boom, particularly during the 2007–2008 period when private intermediaries essentially pulled out of the market. Gabriel and Rosenthal concluded that the government takeover of the GSEs was effective in providing liquidity to the mortgage market during the financial crisis.

Ghent et al. (2013; henceforth GH-MO) estimated the impact of the affordable housing goals on subprime mortgage volume, pricing, and performance. The focus on subprime loans is key to answering the question “did GSE or CRA affordable lending goals contribute to the financial crisis?” since it was the subprime market and not the prime market that imploded. Subprime loans typically cannot be directly purchased by the GSEs since they are nonconforming loans and the majority of subprime loans are not purchased by CRA-eligible entities, so GH-MO focused on the holdings of PLMBS by the GSEs and CRA-eligible depository institutions that are made up of subprime loans.

GH-MO found that while 70% of mortgages in their sample satisfy the affordable housing goals, none of the PLMBS pools that they examined were CRA-qualified. This is because there are very strict guidelines for MBSs to satisfy CRA goals (only loans from a CRA-eligible institution's assessment area count toward the goal). On the other hand, if a GSE purchases a PLMBS and only 20% of the loans in this security satisfy GSE goals, the GSE can count this 20% toward its affordable housing goals. But it is also questionable that the GSEs were primarily buying the PLMBS to satisfy the borrower-related affordable housing goals since the average borrower income to

area median income ratio was 1.73. It appears that the CRA-eligible institutions and the GSEs were buying PLMBSs only for investment purposes.

GH-MO found no significant impact of affordable housing goals (via either CRA or the GSEs) on subprime mortgage volume, pricing, or performance (90 + days delinquent or foreclosures within 2 years of origination). They concluded that it is still possible that the GSEs affected the subprime market by purchasing large numbers of PLMBS since this may have increased credit supply that was used to buy even riskier mortgages.

In summary, a few studies find significant effects of the affordable housing goals on the homeownership rate (An et al., 2007; Gabriel and Rosentha, 2009). But generally, there is little evidence of a significant impact of the GSE and CRA affordable housing goals on the homeownership rate, loan volume, or loan performance. This is the case for both prime and subprime loans. The latter result found by Ghent et al. (2013) is the strongest evidence that the affordable housing goals had no impact on the financial crisis since this was fueled by the poor performance of subprime loans.

14.4.4.4 The GSE impact on mortgage rates

The implicit guarantee of GSE-securitized loans amounts to a subsidy to the GSEs. A question to be answered is “what portion, if any, of this subsidy is passed on to borrowers in terms of lower interest rates?” The GSEs only purchase loans below a specific loan amount and with specific characteristics related to LTV ratio, debt-to-income ratio, credit history, and level of documentation. These are referred to as conforming loans. Loans for amounts above this threshold are referred to as jumbo loans. A reasonably large literature has developed that attempts to estimate the jumbo-conforming spread as a means for measuring the benefits of the GSEs in terms of lower interest rates. The link, then, to homeownership rates is the impact that the lower interest rate has on the likelihood of owning a home.

In an analysis of the literature to date and from his own results based on the commonly used Federal Housing Finance Board’s Mortgage Interest Rate Survey (MIRS), McKenzie (2002) found that the range of estimates of the jumbo-conforming interest rate differential is 20–25 basis points. As Kaufman (2014) notes, though, these studies are subject to selection and sorting bias. Hence, we focus on more recent studies that have attempted to control for these biases.

Ambrose et al. (2004) used a data set that makes up for some of the disadvantages of the MIRS. The data include 26,179 conventional fixed-rate mortgages made between January 1995 and December 1997 by a national lender that includes FICO credit scores that can be used along with loan size to separate loans into conforming and nonconforming loans and into jumbo and nonjumbo loans. They estimated that the interest rate differential between conforming GSE loans and jumbo loans is 24 basis points; 9 basis points are due to the conforming–nonconforming differential and 15 basis

points are due to the conforming loan limit. Ambrose et al. interpreted the former as the pass-through from the GSE subsidy.

In an effort to measure the GSE funding advantage, [Passmore et al. \(2005\)](#) developed a model of interest rate determination for jumbo and conforming spreads that is a function of the differential costs of funding these types of mortgages. They showed that this is not only a function of the GSE funding advantage but also the level of mortgage demand and core deposits. The authors use loan-level data from April 1997 to May 2003 from the MIRS, limiting the sample to about 1 million loans between \$50,000 and twice the conforming loan limit yields. The results show that the jumbo-conforming spread is 15–18 basis points of which 7 basis points can be attributed to the GSE subsidy.

[Sherlund \(2008\)](#) augmented the MIRS data from January 1993 to June 2007 with zip code-level demographic information from the 2000 census. He then used a semiparametric approach based on loan size, LTV, and zip code to control for unobserved differences in borrower and market characteristics and expected house volatility. Sherlund obtained estimates of the jumbo-conforming spread of 13–24 basis points. The lower bound comes from controlling for the endogeneity of loan-type choice by including a sample selection bias correction term.

[Kaufman \(2014\)](#) estimates the impact of GSE purchase eligibility on mortgage interest rates, loan delinquency rates, and mortgage contract features. To control for selection bias and pecuniary market externalities, the author develops an RD design based on the GSE conforming mortgage amount limit. Furthermore, he controls for borrower sorting around the limit using appraisal values as an instrument for mortgage values. The idea is that while mortgage values can be manipulated to meet the conforming limit, appraisal values are not manipulated in this manner.

The sample consists of 14.9 million first-lien, non-FHA, non-VA-insured loans for owner-occupied single-family homes for 2003–2007 from Lender Processing Services Applied Analytics, Inc. Results indicate that GSE purchase eligibility lowers mortgage interest rates by 8–12 basis points. GSE purchase eligibility does not impact default or foreclosure rates or the likelihood of prepayment penalties, negative amortization, interest-only loans, balloon loans, or debt-to-income ratios. Kaufman points out that the fact that default and foreclosure rates are unaffected means that the interest rate impact is fully attributable to the GSE discount and not to less risky borrowers receiving a lower interest rate. The limitation of this approach is that these impacts apply only to borrowers with loan amounts near the conforming loan limit. These tend to be more affluent households and hence the results do not apply to households with moderate or low incomes and/or wealth.

The results from these studies indicate that the GSE subsidy has led to a reduction in interest rates of around 10–20 basis points. The next step is to determine if this has affected the homeownership rate. Note that changes in the homeownership rate are primarily due to first-time homeowners. So the focus needs to be on the interest rate

sensitivity of potential first-time homeowners. Existing homeowners' reaction to interest rates will be on the intensive, not the extensive, margin. That is, lower interest rates can result in cash-out refinancing that is used for home improvements.

Painter and Redfern (2002) investigated the impact of interest rates on both short-run and long-run homeownership rates. Their results confirm the previous findings in the literature that there is little evidence of a significant impact of mortgage rates on short-run or long-run homeownership rates. Hence, it is not likely that the GSE mortgage interest rate subsidy had a significant impact on the homeownership rate.

14.4.4.5 The government role in the mortgage default and foreclosure process

Mortgage default and foreclosure can be seen as the antithesis of homeownership. That is, it typically results in the movement from a state of homeownership to one of nonhomeownership. The drop in the homeownership rate could be due to the fact that the foreclosed properties either were vacant for sustained periods of time, were rented, or were demolished. Furthermore, because of the negative impact on credit scores, households in foreclosed properties are effectively barred from buying another house for many years.⁴¹

To set the stage for the analysis of the government role in the mortgage default and foreclosure process, we first provide information on historical foreclosure rates dating back to 1925 and compare them to the homeownership rate. We then detail a very successful early mortgage modification program, the HOLC. We then focus on the role that borrower-friendly state foreclosure laws have had on foreclosure rates.

14.4.4.5.1 History of foreclosures and the homeownership rate

Historical data are sparse since the federal government does not keep official statistics on foreclosures. Figure 14.8 plots foreclosure rates for 1925–1979. This is based on two series from Snowden (2006): nonfarm foreclosures for 1925–1969 and foreclosures on VA loans for 1950–1979. The foreclosure rate rose steadily during the latter half of the 1920s and reached historical highs of more than 13 per 1000 mortgages in the early 1930s before exhibiting an equally strong decline in the rest of the 1930s. One reason for this strong decline was the implementation of the HOLC (see below). The two series shown in Figure 14.8 line up quite well and both show another spike around 1965 of at least 5 foreclosures per 1000 mortgages.

There was a 5 percentage point drop in the homeownership rate between 1930 and 1940, whereas the foreclosure rate peaked in 1932 and declined steadily until 1940. In fact, the correlation between the nonfarm foreclosure series and the homeownership rate is -0.28 . But given that the homeownership rate is only reported once every 10 years through 1940, it might well be that the decline started prior to 1930, and hence, the actual

⁴¹ Molloy and Shan (2013) found that most households are renters 2 years after a foreclosure start.

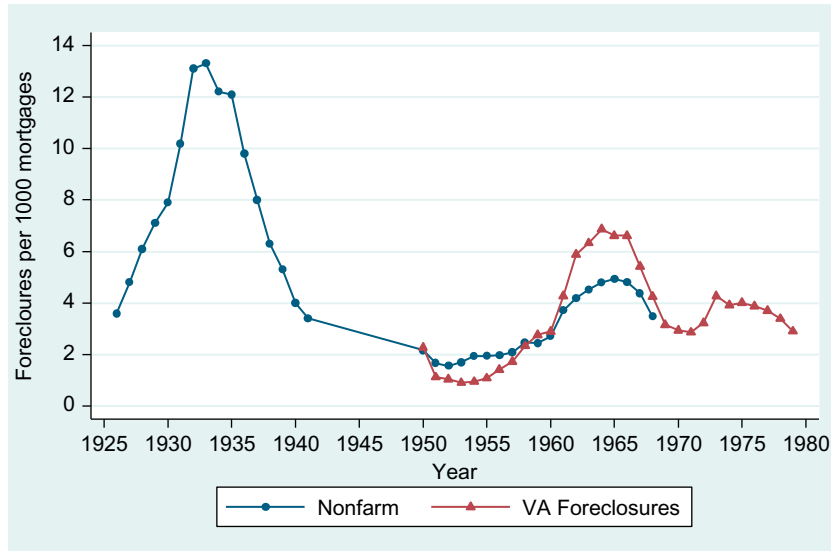


Figure 14.8 Annual US foreclosure rates: 1925–1979. Source: Snowden (2006), Series Dc1255–1270.

correlation between the homeownership and foreclosure rates might well be even stronger than this estimate value based on limited data.

Figure 14.9 plots quarterly foreclosure rates starting in 2003 based on data from Equifax. While this series may not be completely compatible with the earlier data, what is interesting is that the national rate peaked in the second quarter of 2009 at 2.4 foreclosures per 1000 mortgages, which is significantly lower than what was experienced in the Great Depression. Of course, what made this a crisis was the significant heterogeneity across states; for example, the rate peaked at close to 9 foreclosures per 1000 mortgages in the second quarter of 2009 in Nevada.

There was a decline in the US homeownership rate from a high of 69.0% in 2004 to 65.1%. This coincides with the rise in foreclosures during this period. Hartley (2010), Mian et al. (2011), and Anenberg and Kung (2014) show that the increase in foreclosures during this period led to an increase in unsold housing inventory, which is consistent with this drop in homeownership.⁴²

14.4.4.5.2 Home Owners' Loan Corporation

In 1933, the Home Owners' Loan Corporation Act established HOLC to refinance distressed mortgages. The goal of HOLC was to keep homeowners in their homes and stabilize the homeownership rate. Applications were accepted between June 1933

⁴² To get a causal impact of foreclosure on homeownership, one could do the same analysis as in Mian et al. (2011) using the homeownership rate instead of unsold housing inventory as the dependent variable.

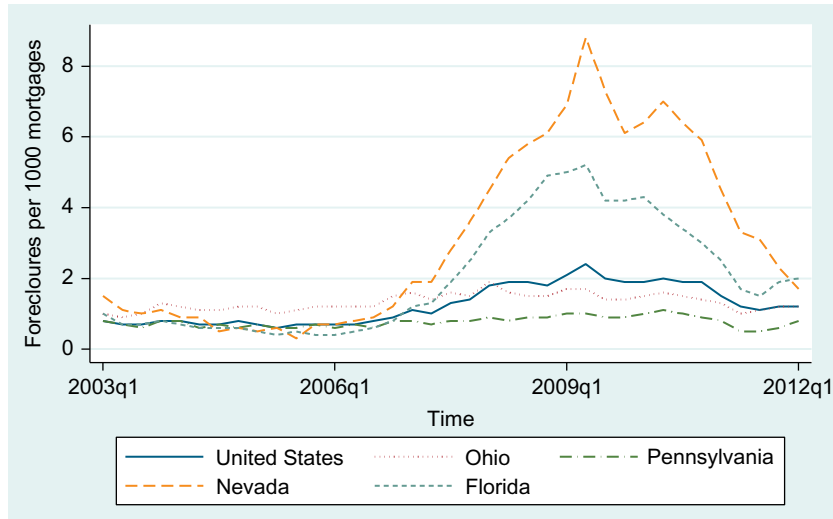


Figure 14.9 Quarterly US foreclosure rate: 2003q1–2012q1. *Source: FRBNY Consumer Credit Panel/Equifax. Note: Based on the population with a credit report.*

and November 1934 and in May and June 1935. After that, the loan portfolio gradually declined until the program was ultimately ended in 1951 and existing loans were sold to private lenders.

Rose (2011) noted that refinancing was requested for 40% of all mortgaged 1–4 family nonfarm units in the United States and approximately 1 million loans for more than \$3 billion were made. HOLC purchased loans from the lender using bonds that were essentially the same as US T-bills and then made new 15-year self-amortizing loans to the borrowers at a 5% interest rate. These loans had no prepayment fee and borrowers could make only interest payments for the first 3 years (Rose, 2011). This represented an important shift in loan structure, as existing mortgages tended to be of short duration, were not self-amortizing, and required balloon payments upon maturity, which tended to be refinanced. This process was essentially halted by the Great Depression and foreclosure rates skyrocketed (see Figure 14.8).

HOLC loans did not exceed 80% of the property appraisal. If the value of the initial loan was greater than 80%, HOLC only accepted the loan if the lender agreed to be paid the 80% figure, hence taking a loss on the loan. Of course, this “loss” is mitigated since the lender does not have to incur the costs associated with foreclosure. Rose reported that approximately 18% of loan applications were rejected because the lender refused to take the principal reduction needed to make the 80% threshold. Using data from actual HOLC loans from New York, New Jersey, and Connecticut, Rose found that the appraisals were inflated as a motivation to get lenders to participate in the HOLC program. This means that the reductions provided to borrowers were lower. Rose estimated

that principal reductions would have been 16% on average versus 8% if appraisal values had been set to actual market values. Rose argued that HOLC was more concerned about stabilizing the housing market versus getting the best deal for each homeowner with the rationale that a stabilized market and the general economic recovery would lead to lower overall foreclosure rates.

The high participation rate under HOLC can be attributed to a number of factors. First was the generous terms given to the lenders. Second, the new structure of the mortgages provided by HOLC was a real benefit to borrowers, and third, underwriting standards were higher in the 1920s and 1930s such that LTVs were significantly lower then, making the average write-down that a lender had to accept to participate in HOLC much less than under current loan modification programs (Rose, 2011).

14.4.4.5.3 Borrower-friendly foreclosure laws

There was substantial variation in foreclosure rates across states as displayed in Figure 14.9. Part of this was due to differences in the population characteristics and incomes across states, but part was due to the differences in foreclosure laws. Three important distinctions in foreclosure laws pertain to whether recourse is allowed, whether the foreclosure process is judicial (vs. nonjudicial), and whether a redemption period is allowed. All of these can be seen as borrower-friendly characteristics. So the question is to what extent do these borrower-friendly laws affect homeownership rates? That is, do borrower-friendly laws help homeowners to stay in their homes and avoid foreclosure or do they facilitate foreclosure and actually reduce homeownership rates? Furthermore, to the extent that these laws increase the cost to lenders of going through the foreclosure process, will lenders be less likely to originate loans for the marginal borrower?

Typically, after the third month of delinquent payments, the lender will begin the foreclosure process. All states allow for a judicial foreclosure process, while 29 states also allow for a nonjudicial process known as “power of sale.” In the former case, the foreclosure process occurs through the courts. This is initiated with the lender filing suit in court that details the debt owed by the homeowner and the reason for foreclosure. A notice or “lis pendens” is sent to the owner that demands payment of the debt owed. Typically, the owner has 30 days to pay off the debt. If not, the lender can request a sale of the property by auction.

In the case of a nonjudicial foreclosure process, the lender sends a “notice of default” to the owner and to the local jurisdictional authority. If the borrower does not respond by repaying what he owes from delinquency, this will result in the initiation of the auction process with the filing of a “notice of sale” by the lender. In either a judicial or a nonjudicial foreclosure process, the auction can take place as soon as 2–3 months after a request to do so is made by the lender. The actual time from the beginning of the foreclosure process to the date of sale will vary from state to state and also depends on the type

of foreclosure process. Typically, this is significantly longer for judicial versus nonjudicial states (Pence, 2006; Pennington-Cross, 2010). For example, this process typically takes 18 months in New York, a judicial state (Schuetz et al., 2008). During this time, the occupant can live in the house without any payments to its owner.

The actual foreclosure of the property occurs when the property is sold at auction (about 20% of the time) or, if unsold, is transferred to the lender. In the latter case, the property is said to be “real estate-owned” (REO) since it is transferred to the lender’s REO department. Once ownership is transferred to a third party via a successful auction or to the lender, the previous owner becomes a tenant of the property. The new owner must follow state procedures to evict the previous owner from the unit.

Some states allow the borrower an opportunity to regain the rights to the property even after the sale date during what is known as the redemption period.⁴³ To do so, the borrower must repay the auction price and foreclosure expenses during this period, which can last up to 18 months. Typically, a redemption period is only available under a judicial foreclosure. Of course, the ability of the borrower to repay these costs is very unlikely unless she “wins the lottery.”

Another factor that affects foreclosure is whether mortgages are considered to be recourse debt. If so, upon default, the lender can lay claim to the debtor’s assets to cover the difference between the value of the mortgage and the fair market value of the foreclosed property. This process is known as a deficiency judgment. The process of filing a deficiency judgment and actually collecting from the borrower differs across states. Furthermore, the borrower can file for Chapter 7 bankruptcy to protect her assets. Table 1 in Ghent and Kudlyak (2011) provides information on the extent of recourse in each state and classifies each state as a recourse or nonrecourse state (11 are classified as nonrecourse).

Occupants can avoid a foreclosure by paying back the delinquent debt or by negotiating with the lender to allow for what is known as a short sale. This is a purchase of the property by a third party for a price that is less than the outstanding principal on the mortgage. Lenders will often agree to a short sale to avoid the costs of holding a property as REO and then having to sell it themselves on the open market. This negotiation process can also result in the owner having to pay the back all or part of the difference between what is owed on the mortgage and the sales price. A successful short sale also results in lower legal costs for the lender.

A necessary condition for foreclosure is that the property is worth less than the mortgage balance plus selling costs (otherwise, the owner can sell the property and payback the mortgage debt). The “double-trigger theory of default” posits that this is not sufficient for foreclosure. What is also required is that the household experiences a major adverse life

⁴³ As of January 2011, foreclosure redemption is only possible in 25 states http://www.ehow.com/info_7867003_redemption-foreclosure.html#ixzz2wnAL21Ev.

event such as death, the onset of a severe health condition, job loss, or divorce that reduces desired housing consumption (Foote et al., 2008). Another factor that plays a role in the foreclosure decision is future price expectations. If prices are expected to rise, a homeowner will be less likely to default.

Using 1994 and 1995 HMDA data and applying a border fixed effects approach to MSAs that cross state lines, Pence (2006) found that the size of loans is lower in judicial states; that is, borrower-friendly laws can reduce the supply of mortgage credit as this increases costs to lenders. This has the potential for adversely affecting homeownership. Collins et al. (2011) used a similar border fixed effects approach to loans that were 60+ days delinquent as of January 2007 and their status 15 months later (cured, foreclosed, or still delinquent). Their results show that none of the three state foreclosure policies they evaluated (judicial foreclosure proceedings, statutory rights of redemption, and statewide foreclosure prevention initiatives) had a significant impact on the loan status after 15 months. Using 2005 and 2006 HMDA data, Curtis (2014) used the same border fixed effects approach to provide evidence that borrower-friendly laws result in a smaller sub-prime market share.

Gerardi et al. (2013) found that while foreclosure rates were lower in judicial states in the short run, this was not the case in the long run. The result is that being in a judicial state just lengthened the foreclosure process. They found that this was also the case for a “right-to-cure” law that was enacted in Massachusetts in 2008.

Using monthly data from the Lender Processing Services Applied Analytics database on nearly 3 million prime and nonprime private securitized loans, portfolio loans, and GSE loans that were originated between August 1997 and December 2008, Ghent and Kudlyak (2011) found that while both unconditional and conditional default rates do not differ across recourse and nonrecourse states, default rates are higher in the latter states for underwater borrowers. This implies that when housing markets are in decline and borrowers are more likely to be underwater, the borrower-friendly law that prohibits recourse can increase default rates. Furthermore, when borrowers do default in recourse states, they are more likely to cure and more likely to end up in a friendly foreclosure. These results point to higher foreclosure rates in nonrecourse states. When looking at the type of loan, Ghent and Kudlyak found that the impact of recourse on default only holds for privately held mortgages and not those held by the GSEs.

The Bankruptcy Abuse Prevention and Consumer Protection Act (BAR) was an attempt to reduce bankruptcy abuse by introducing means testing for Chapter 7 filing and hence can be seen as a consumer-hostile law. Prior to this, individuals could file for Chapter 7 bankruptcy and clear their unsecured debts (such as credit card debts) to allow them to pay their mortgages. Morgan et al. (2012) tested to see if BAR led to increased foreclosures in states with higher homestead exemptions. This is because prior to BAR, higher home equity exemptions meant that filers had more funds to pay their mortgages. They also hypothesized that this would impact foreclosures of

subprime mortgages and not prime mortgages since holders of prime mortgages are unlikely to need to file for Chapter 7 bankruptcy. Using a panel of quarterly state-level data from 1998:1 to 2007:3, the authors showed that BAR did increase foreclosure rates for borrowers with subprime mortgages in states with higher homestead exemptions. There was no impact on foreclosure rates for borrowers with prime mortgages. This is evidence that borrower-unfriendly laws can have adverse effects on foreclosure rates.

Overall, these studies do not provide consistently strong evidence that borrower-friendly laws help borrowers in reducing foreclosure and ultimately homeownership rates. At best, these results are inconclusive. More research is needed to clarify this relationship.

14.4.5 The government response to the housing crisis

In response to the recent financial and housing crisis, the federal government enacted a number of policies. The first response that potentially affected the homeownership rate was the Housing and Economic Recovery Act (HERA) of 2008 that was enacted on 30 July 2008.⁴⁴ Among other things, this act merged the Federal Housing Finance Board and OFHEO into the Federal Housing Finance Agency (FHFA). One of its first significant moves was to put Fannie Mae and Freddie Mac under the conservatorship of the FHFA. HERA also includes the FHA Modernization Act, the Housing Assistance Tax Act (HATA), and the HOPE for Homeowners (H4H) Act.

The FHA Modernization Act (1) increased the FHA loan limit from 95% to 110% of area median home prices up to 150% of the GSE conforming loan limit, effective 1 January 2009; (2) required a down payment of at least 3.5% for any FHA loan; (3) placed a 12-month moratorium on HUD's implementation of risk-based premiums; (4) prohibited seller-financed down payments; and (5) allowed down payment assistance from family members.

Due to the more than doubling of the loan limit under the FHA Modernization Act, FHA's market share of single-family home purchase loans jumped from a low of 4% in 2007, to almost 25% in the third quarter of 2008, to a high of 32% in the second quarter of 2010 and remained around 25% in 2011 and 2012 (HUD data). While the FHA was able to help fill the gap left by the exit of private institutions from the mortgage market, the large increase in volume, particularly of high LTV loans in 2008 and 2009, led to high delinquency rates and losses that may (still) require a taxpayer bailout. Despite these problems, [Van Order and Yezer \(2014\)](#) stated that "The Federal Housing Administration (FHA) deserves considerable credit for helping support the housing market during the recent financial crisis by increasing its own market share."

HATA established a first-time home buyer tax credit for purchases between 9 April 2008, and 1 July 2009. The tax credit was 10% of the purchase price up to a credit of

⁴⁴ See http://en.wikipedia.org/wiki/Housing_and_Economic_Recovery_Act_of_2008.

\$7500. Those receiving the credit had to repay it over 15 years. Also, the credit was phased out for households filing joint returns with incomes over \$150,000 and other households with incomes over \$75,000. HATA also provided assistance for the redevelopment of abandoned and foreclosed properties. As part of the American Recovery and Reinvestment Act (ARRA) of 2009, the payback of the credit was set at \$8000 for homes purchased between 1 January 2009, and 30 November 2009, and there was no repayment if the house was not sold within 3 years. Finally, the Worker, Homeownership, and Business Assistance Act (Assistance Act) of 2009 extended the time frame in which homebuyers could claim the ARRA version of the credit to 30 April 2010. The GAO reported that as of July 2010, approximately 1 million first-time homebuyers claimed \$7.3 billion in interest-free loans under HATA, whereas 16 million first-time homebuyers claimed approximately \$23 billion in tax credits under the ARRA and Assistance Act provisions.

Baker (2012) claimed that the initiation of the first-time homebuyer credit led to an increase in sales and prices and was followed by a drop in sales and prices once it ended. Hence, the initial increase in price means that much of the credit was passed on to buyers. But the evidence in the paper for a causal impact is tenuous as it does not rely on a quasi-experimental framework. Still, Baker was likely correct in pointing out that the credit only shifted home purchases forward since the credit recipients would likely have ended up buying at a later point and hence had little impact on the homeownership rate.

The H4H Act authorized the FHA to insure up to \$300 billion of 30-year fixed-rate refinance loans up to 96.5% of LTV for delinquent borrowers and to 90% for borrowers with a low credit score for mortgages that were made on or before 1 January 2008. The federal government was to receive 50% of any appreciation in value of the house. In return, lenders received insurance under FHA. The goal was to modify 400,000 loans, but only 64 were actually completed by August 2010. The lack of response is attributed to the minimal incentives for lenders (Rose, 2011).

The second federal response to the housing crisis was the Emergency Economic Stabilization Act (EESA) of 2008 that was passed on 3 October 2008. EESA authorized the Treasury to establish the Troubled Assets Relief Program (TARP) to purchase troubled assets. As part of TARP, the Treasury established Making Home Affordable (MHA). The purpose of MHA was to modify loans of distressed homeowners and to prevent some foreclosures. MHA established the Home Affordable Modification Program (HAMP) and the Home Affordable Refinance Program (HARP). Both will continue until 31 December 2015.

HAMP set a target of modifying the loans of 3–4 million distressed homeowners (US GAO Month in Review July 2009). To be eligible, the owner had to occupy the unit as his or her primary residence. The unit had to be single family with an unpaid mortgage balance no greater than \$729,750 that was originated on or before 1 January 2009.

HAMP was based on the monthly first-lien mortgage payment (principal, interest, property taxes, and homeowner's insurance payments) to gross monthly income ratio (PI) rather than on LTV. The goal was to reduce PI to 31%, so only homeowners with a PI of greater than 31% were eligible under HAMP. The lender first offers a trial modification that becomes permanent if the borrower makes the new payments for approximately 6 months.

A key component of HAMP is the net present value (NPV) calculation that determines if the loan modification is beneficial as compared to foreclosure. A positive NPV requires the lender to curtail the foreclosure process and offer a HAMP modification. A negative NPV leaves it up to the lender's discretion to continue the foreclosure process or offer a HAMP modification.

Note that homeowners do not have to be delinquent to be eligible for HAMP. It is stated that borrowers are eligible if they are 60 days or more delinquent or are facing economic hardship and are determined to be in imminent default.⁴⁵ In fact, there are additional monetary incentives to lenders and investors to modify loans where the borrower is current on mortgage payments.

Under HAMP, lenders first reduce the interest rate, then extend the term of the loan to a maximum of 40 years, and finally set up a balloon payment at the end of the loan term as a means for reducing PI to 31%—no principal write-down is required. As compensation, the lender receives a one-time payment of \$1000. The loan servicer could receive up to \$1000 a year for three years based on the payment history of the borrower.

HARP was established to help nondistressed homeowners (current in mortgage payments) who experienced negative equity (or close to negative equity) to refinance their loans. In particular, HARP allowed homeowners with mortgages owned or guaranteed by Freddie Mac or Fannie Mae that were purchased before 31 May 2009, and with an LTV greater than 80% and up to 125% to refinance without having to purchase private mortgage insurance if they did not already have it. The refinance must result in lower monthly payments or be a more stable product (i.e., moving from an ARM to a 30-year fixed-rate mortgage). The so-called HARP 2.0 was a modification of HARP that allowed for refinancing regardless of the LTV. Plus lenders that refinanced under HARP were not held responsible for fraud committed on the original loan.⁴⁶

Agarwal et al. (2012a) evaluated the impact of HAMP on the number of loan modifications, the characteristics of these modifications, and the likelihood of default. They constructed one control group from loans that are investor-owned (not owner-occupied) since these are not eligible for HAMP. Using pre-HAMP data, they showed that investors and owner-occupiers have similar LTVs, interest rates, modification rates, and delinquency rates though investors have slightly higher FICO scores and higher foreclosure

⁴⁵ For example, see http://www.freddiemac.com/singlefamily/service/hmp_eligibility.html.

⁴⁶ See http://en.wikipedia.org/wiki/Home_Affordable_Refinance_Program.

rates than owner-occupiers.⁴⁷ A second control group is based on an RD design that focuses on the maximum loan balance that is HAMP-eligible, \$729,750. The drawback is that there are few owners in financial distress with mortgage balances just below \$729,750 that receive loan modifications under HAMP, and hence, the results might not be generalizable.

Based on the first control group of investors, the authors estimate the short-term impact of HAMP led to an additional 1.2 million modifications and reduced the number of foreclosures by 800,000 as of December 2012, both of which are well below the program goals. Using the second control group produces results that are consistent with the first identification strategy.

Finally, Agrawal et al. found that the participation rate in HAMP of a few large loan servicers (accounting for 75% of the loans) was half that of other servicers. Across all servicers, these participation rates were similar to pre-HAMP mortgage renegotiation rates and can be explained by an organizational design that was conducive (or not) for undertaking loan modifications. Hence, one reason for the low take-up of HAMP was that a large majority of the loans were serviced by firms that were not structured to make loan modifications. The authors estimate that if the loan modification rate of the low-participation rate servicers was doubled so as to be the same as that of the high-participation-rate servicers, the number of modifications under HAMP would increase by 70%.

Hembre (2014) used a computable GE model of the mortgage default decision to calculate the benefits of HAMP. Identification of the model parameters comes from matching to sample moments of the data. Hembre used the parameter estimates to simulate counterfactual scenarios such as the absence of HAMP. He used data on the 1.1 million households that received a permanent HAMP mortgage modification. The average reduction in annual mortgage payment was \$9900 that was brought about by a nearly 4% drop in the interest rate, by an extension of the mortgage term by 4.5 years, and by a 6% reduction in principal balance. The average difference between the loan and the house value is \$54,000 or 39% for current HAMP participants as of June 2013. Matching to a second data set with information on second mortgages (20% of the sample holds a second mortgage) and to Zillow real estate data reduces the sample size to 5629 (this huge reduction in sample size is a drawback of the paper).

Hembre found that HAMP prevented 515,354 defaults as of June 2013 and a 5-year projection shows HAMP preventing 505,803 defaults. Current HAMP costs are calculated to be \$9.5 billion and expected 5-year costs are \$20.8 billion or \$41,600 per prevented foreclosure. Based on estimates of foreclosure externalities from Campbell et al.

⁴⁷ Despite what seems like a reasonable matching on observables, this is a questionable choice of control group given that investors as generally riskier borrowers than owner-occupants (Mayer et al., 2009). The key here might be limiting loans to those originated by main banking institutions.

(2011) and [Harding et al. \(2012\)](#), Hembre estimated the benefit of one less foreclosure to be \$16,000. This amounts to a total social benefit of \$8 billion and hence a net loss of \$12.7 billion. Hembre then simulated the impact when HAMP modifies the loan so that payments are 25% and 38% of income. The former raises program costs and cost per prevented foreclosure by 29% and 52%, respectively. The latter lowers program costs and cost per prevented foreclosure by 45% and 23%, respectively. The breakeven level is the reduction of payment level to 52% of income for which only 25% of the existing HAMP participants would be eligible.

In a theoretical analysis, [Mulligan \(2010\)](#) found that one reason why HAMP has been ineffective is its incentive structure that targets the PI ratio and the uncertainty it has created for borrowers have made it “impossible to both write down principal and offer modification to a wide range of borrowers.”

The federal government implemented numerous policies to curtail the financial and housing crisis. In the housing market, many policies focused on reducing the foreclosure rate. These included the H4H, the HAMP, and the HARP. Overall, the evidence shows that the government effort at curtailing the foreclosure crisis was generally ineffective. The lone exception may be policies that led to a large increase in the FHA’s market share at the time that private institutions exited the market.

One reason that has been made for why programs such as HAMP and HARP were unsuccessful is that securitization, because of its complexity, has made modification of loans more difficult.⁴⁸ [Adelino et al. \(2009\)](#) use an IV approach to provide evidence that securitization actually increases the likelihood that mortgages will be modified and decreases the chance they will be foreclosed on by mortgage servicers. The authors argue, instead, that owners of mortgage-backed securities are deterred from loan renegotiations because they are concerned that borrowers will still default after costly modifications and because delinquent borrowers will self-cure even without the modifications.

14.4.6 Conclusion

Governments in the United States have attempted to affect the homeownership rate through a wide range of policies. We have focused on federal and state policies that are related to the mortgage market, particularly the structure, volume, and quality of mortgages. We began in the 1930s with the federal response to the depressed economic and housing conditions that characterized the Great Depression and continued through the federal response to the recent financial crisis. One takeaway is that despite the continued focus on homeownership as the “American Dream,” most of the government policies pursued have had little effect on the homeownership rate, particularly in the long run. In fact, some of the largest federal government involvements in the mortgage market

⁴⁸ For example, see <http://www.jdsupra.com/legalnews/securitization-and-loan-modification-88901/>.

such as the NHA of 1934 and the VA Mortgage program were primarily intended as economic stimulus policies.

The homeownership rate was pretty stable between 1880 and 1930, and despite the large government involvement in the mortgage market in the 1930s, the homeownership rate fell by 5 percentage points. Despite this drop, the implementation of the FHA, Fannie Mae, and programs such as HOLC likely prevented the homeownership rate from falling even further.

The following 20 years saw a significant increase in the homeownership rate, and it is during this time when federal and state policies had a large impact. [Fetter \(2013b\)](#) made a strong case that the imposition of rent control during WWII led to a significant increase in the homeownership rate, and [Fetter \(2013a\)](#) and [Chambers et al. \(2013\)](#) provided evidence that the introduction of the GI Bill and the VA Mortgage program significantly impacted the structure of mortgage finance and the homeownership rate. Along with the FHA mortgage insurance program, the VA program offered longer loan terms with higher LTVs. The latter characteristic allowed otherwise wealth-constrained households to purchase homes with small down payments.

[Chambers et al. \(2009\)](#) found some evidence that changes in mortgage finance such as piggyback loans, first introduced by the GSEs in the late 1990s, affected the increase in the homeownership rate between 1994 and 2005. But this was a short-term effect as this financing just shifted forward the age at which some households first purchased a home. Furthermore, it led to many households purchasing homes who were unable to make mortgage payments and who soon defaulted and then lost their homes to foreclosure. The federal role in this short-term increase in the homeownership rate likely traces back to the deregulation of the mortgage market that began in the early 1980s that laid the groundwork for the dramatic rise in subprime loans made by large and national in scope nondepository institutions.

A significant indirect government presence in the mortgage market is captured by the GSEs Fannie Mae and Freddie Mac. While, historically, they had a small role in the mortgage market, their share of outstanding mortgage debt started to increase dramatically in the 1980s. Concurrent with this increased GSE activity, and together with the passing of CRA in 1977, the 1992 GSE Act mandated steps to increase loans to low-income households and in low-income areas. Many viewed these affordable housing goals as leading to a lowering of mortgage standards and a primary reason for the recent housing crisis. But there is little evidence of a significant impact of the affordable housing goals on the homeownership rate (as well as loan volume and performance), and it points to the conclusion that, in this respect, the GSEs and CRA probably played a minor role in the recent mortgage crisis.

Part of the reason that the GSEs had little role in the housing crisis is that there is strong evidence of considerable GSE crowd out during the market boom of 2003–2006. Furthermore, given that there was very little crowd out in 2007–2008,

[Gabriel and Rosenthal \(2010\)](#) concluded that the government takeover of the GSEs was effective in providing liquidity to the mortgage market during the financial crisis. This is further evidence that the federal role in the housing market is best justified during significantly down markets (as during the Great Depression and Great Recession).

This countercyclical role of federal intervention in the mortgage market also extends to the FHA. Its market share of single-family mortgage originations was around 3% between 2004 and 2007 when the market was booming and was around 25% starting in 2009 when it was in a major downturn. [Van Order and Yezer \(2014\)](#) stated that this is consistent with the purpose of FHA when it was founded in 1934 to spur the housing market during the Great Depression. Now that the housing market is on more firm footing, Van Order and Yezer advocated that the FHA loan limits should be reduced to historic levels so that it can focus, again, on insuring mortgages to first-time low-income and minority households.

Another reason for the minimal role of the GSEs in the financial crisis was that it was not directly linked to problems in the subprime market. Since subprime mortgages are, for the most part, not conforming, they cannot be directly purchased by the GSEs. The GSEs did purchase subprime PLMBS, but [Ghent et al. \(2013\)](#) showed that the primary reason for doing so was not to meet the affordable housing goals. That said, the GSEs had a large stake in the subprime PLMBS market, and while they only purchased AAA tranches, they tended to invest in riskier securities ([Adelino et al., 2014](#)). This may have increased credit supply that was used to buy even riskier mortgages. However, it could be that there was substantial GSE crowd out in the PLMBS market and the lowering of mortgage quality would have happened even without GSE involvement.

States have enacted borrower-friendly foreclosure laws in the form of requiring a judicial process for foreclosures, redemption periods, and nonrecourse upon foreclosure. The literature provides little evidence that these laws significantly reduced foreclosures or ultimately homeownership rates. That said, one of the more successful federal housing policies was the HOLC that was established in 1933 to refinance distressed mortgages. The popularity of HOLC (40% of all mortgaged 1–4 family nonfarm units requested refinancing under HOLC) was attributed to the generous terms given to the lenders and the fact that the new structure of the mortgages provided by HOLC was a real benefit to borrowers. This success is in stark contrast to the government response to the recent housing crisis, particularly policies aimed at curtailing the foreclosure crisis, which have been shown to be ineffective. Policymakers would have benefited from taking a closer look at HOLC when designing these recent programs.

14.5. CONCLUSION

This chapter illustrates the great diversity of government housing policies even within a single country. It describes the rationales for two of the largest types of housing policies in

the United States, namely, low-income rental assistance and policies to promote homeownership through interventions in mortgage markets, and the nature of the largest programs of each type, the empirical evidence on their effects, and the data and methods used to obtain them. Our review of the evidence finds large differences in the performance of different programs that have been used to pursue the same goals. It also reveals many important gaps in knowledge. Both are surely true for housing policies in other countries. However, much less is known about them.

REFERENCES

- Abt Associates Inc., Mills, G., Gubits, D., Orr, L., Long, D., Feins, J., Kaul, B., Wood, M., Amy Jones & Associates, Cloudburst Consulting, The QED Group, 2006. *Effects of Housing Vouchers on Welfare Families*. U.S. Department of Housing and Urban Development, Office of Policy Development & Research, Washington, DC.
- Adelino, M., Gerardi, K., Willen, P., 2009. Why don't lenders renegotiate more home mortgages? Redefaults, self-cures, and securitization: Federal Reserve Bank of Boston Public Policy Discussion Paper 09-4.
- Adelino, M., Frame, W.S., Gerardi, K.S., 2014. The Effect of Large Investors on Asset Quality: Evidence from Subprime Mortgage Securities (unpublished mimeo, March 17).
- Agarwal, S., Amromin, G., Ben-David, I., Chomsisengphet, S., Piskorski, T., Seru, A., 2012a. Policy intervention in debt renegotiation: evidence from the Home Affordable Modification Program: NBER Working paper 18311.
- Agarwal, S., Benmelech, E., Bergman, N., Seru, A., 2012b. Did the community reinvestment act (CRA) lead to risky lending?: NBER Working paper 18609.
- Ambrose, B.W., Thibodeau, T.G., 2004. Have the GSE affordable housing goals increased the supply of mortgage credit? *Reg. Sci. Urban Econ.* 34, 263–273.
- Ambrose, B.W., LaCour-Little, M., Sanders, A.B., 2004. The effect of conforming loan status on mortgage yield spreads: a loan level analysis. *Real Estate Econ.* 32, 541–569.
- An, X., Bostic, R.W., Deng, Y., Gabriel, S.A., Green, R.K., Tracey, J., 2007. GSE loan purchases, the FHA and housing outcomes in targeted, low-income neighborhoods [with comments]. In: Burtless, G., Pack, J.R. (Eds.), *Brookings-Wharton Papers on Urban Affairs*. Brookings Institution Press, Washington DC, pp. 205–256.
- Anenberg, E., Kung, E., 2014. Estimates of the size and source of price declines due to nearby foreclosures. *Am. Econ. Rev.* 104 (8), 2527–2551.
- Angrist, J., Pischke, S., 2009. *Mostly Harmless Econometrics: An Empiricist's Companion*. Princeton University Press, Princeton, NJ.
- Anil, B., Sjoquist, D.L., Wallace, S., 2010. The effect of a program-based housing move on employment: HOPE VI in Atlanta. *South. Econ. J.* 77 (1), 138–160.
- Avery, R.B., Brevoort, K.P., 2011. The subprime crisis: is government housing policy to blame?: Federal Reserve Board of Governors Finance and Economics Discussion Series Working paper 2011-36.
- Avery, R.B., Bostic, R.W., Canner, G.B., 2005. Assessing the necessity and efficiency of the community reinvestment act. *Hous. Policy Debate* 16 (1), 143–172.
- Bajari, P., Kahn, M.E., 2005. Estimating housing demand with an application to explaining racial segregation in cities. *J. Bus. Econ. Stat.* 23 (1), 20–33.
- Baker, D., 2012. *First Time Underwater: The Impact of the First-time Homebuyer Tax Credit*. Center for Economic and Policy Research.
- Baum-Snow, N., Ferreira, F., 2015. Causal inference in urban economics. In: Duranton, G., Vernon Henderson, J., Strange, W. (Eds.), *Handbook of Regional and Urban Economics*, vol. 5. North-Holland, Amsterdam.

- Baum-Snow, N., Marion, J., 2009. The effects of low income housing tax credit developments on neighborhoods. *J. Public Econ.* 93 (5–6), 654–666.
- Bayer, P., McMillan, R., 2012. Tiebout sorting and neighborhood stratification. *J. Public Econ.* 96 (11–12), 1129–1143.
- Bayer, P., Ferreira, F., McMillan, R., 2007. A unified framework for measuring preferences for schools and neighborhoods. *J. Polit. Econ.* 115 (4), 588–638.
- Bayer, P., McMillan, R., Murphy, A., Timmons, C., 2011. A dynamic model of demand for houses and neighborhoods: National Bureau of Economic Research WP17250, Cambridge, MA.
- Belsky, E.S., Duda, M., 2002. Asset appreciation, timing of purchases and sales, and returns to low-income homeownership. In: Retsinas, N.P., Belsky, E.S. (Eds.), *Low-Income Homeownership: Examining the Unexamined Goal*. The Brookings Institution Press, Washington, DC, pp. 208–238.
- Benjamin, J.D., Sirmans, C.F., 1987. Who benefits from mortgage revenue bonds? *Natl. Tax J.* 40, 115–120.
- Bhutta, N., 2011. The Community Reinvestment Act and mortgage lending in lower-income neighborhoods. *J. Law Econ.* 54 (4), 953–983.
- Bhutta, N., 2012. GSE activity and mortgage supply in lower-income and minority neighborhoods: the effect of the affordable housing goals. *J. Real Estate Financ. Econ.* 45 (1), 238–261.
- Bhutta, N., Canner, G.B., 2009. “Did the CRA Cause the Mortgage Market Meltdown,” *Community Dividend*, March.
- Bingley, P., Walker, I., 2001. Housing subsidies and work incentives in Great Britain. *Econ. J.* 111 (471), C86–C103.
- Boehm, T.P., Schlottmann, A., 1999. Does home ownership by parents have an economic impact on their children? *J. Hous. Econ.* 8, 217–232.
- Boehm, T.P., Schlottmann, A., 2008. Wealth accumulation and homeownership: evidence for low-income households. *Cityscape* 10 (2), 225–256.
- Bolotnyy, V., 2014. The Government-sponsored enterprises and the mortgage crisis: the role of the affordable housing goals. *Real Estate Econ.* 42 (3), 724–755.
- Bostic, R.W., Gabriel, S., 2006. Do the GSEs matter to low-income housing markets? An assessment of the GSE loan purchase goals on California housing outcomes. *J. Urban Econ.* 59, 458–475.
- Bostic, R.W., Engel, K.C., McCoy, P.A., Anthony Pennington-Cross, A., Wachter, S.M., 2008. State and local anti-predatory lending laws: the effect of legal enforcement mechanisms. *J. Econ. Bus.* 60 (1–2), 47–66.
- Bradbury, K., Downs, A., 1981. *Do Housing Allowances Work?* Brookings Institution, Washington, DC.
- Burge, G.S., 2011. Do tenants capture the benefits from the low-income housing tax credit programs? *Real Estate Econ.* 39 (1), 71–96.
- Buron, L., Nolden, S., Heintzi, K., Stewart, J., 2000. *Assessment of the Economic and Social Characteristics of LIHTC Residents and Neighborhoods: Final Report*. U.S. Department of Housing and Urban Development, Office of Policy Development and Research, Washington, DC.
- Campbell, J., Giglio, S., Pathak, P., 2011. Forced sales and house prices. *Am. Econ. Rev.* 101 (5), 2108–2131.
- Carliner, M.S., 1998. Development of federal homeownership “Policy” *Hous. Policy Debate* 9 (2), 299–321.
- Carlson, D., Haveman, R., Kaplan, T., Wolfe, B., 2012a. Long-term effects of public low-income housing vouchers on neighborhood quality and household composition. *J. Hous. Econ.* 21 (2), 101–120.
- Carlson, D., Haveman, R., Kaplan, T., Wolfe, B., 2012b. Long-term earnings and employment effects of housing voucher receipt. *J. Urban Econ.* 71 (1), 128–150.
- CBO, 2010. *Fannie Mae, Freddie Mac, and the Federal Role in the Secondary Mortgage Market*. CBO, Washington, DC.
- Chambers, M., Garriga, C., Schlagenhauf, D.E., 2009. Accounting for changes in the homeownership rate. *Int. Econ. Rev.* 50 (3), 677–726.
- Chambers, M., Garriga, C., Schlagenhauf, D.E., 2013. *Constructing the Post-War Housing Boom* (unpublished mimeo).
- Collins, J.M., Lam, K., Herbert, C.E., 2011. State mortgage foreclosure policies and lender interventions: impacts on borrower behavior in default. *J. Policy Anal. Manage.* 30 (2), 216–232.

- Coulson, E., Li, H., 2013. Measuring the external benefits of homeownership. *J. Urban Econ.* 77, 57–67.
- Coulson, N.E., Imai, S., Hwang, S.J., 2003. The benefits of owner-occupation in neighborhoods. *J. Hous. Res.* 14 (1), 21–48.
- Currie, J., Yelowitz, A., 2000. Are public housing projects good for kids? *J. Public Econ.* 75 (1), 99–124.
- Curtis, Q., 2014. State foreclosure laws and mortgage origination in the subprime. *J. Real Estate Financ. Econ.* 49, 303–328.
- Demyanyk, Y., Van Hemert, O., 2011. Understanding the subprime mortgage crisis. *Rev. Financ. Stud.* 24 (6), 1854–1880.
- Dietz, R.D., Haurin, D.R., 2003. The social and private micro-level consequences of homeownership. *J. Urban Econ.* 54, 401–450.
- DiPasquale, D., Glaeser, E.L., 1999. Incentives and social capital: are homeowners better citizens? *J. Urban Econ.* 45, 354–384.
- Durning, D., Quigley, J.M., 1985. On the distributional implications of mortgage revenue bonds and creative finance. *Natl. Tax J.* 38 (4), 513–523.
- Early, D.W., 1998. The role of subsidized housing in reducing homelessness: an empirical investigation using micro-data. *J. Policy Anal. Manage.* 17 (4), 687–696.
- Early, D.W., 2004. The determinants of homelessness and the targeting of housing assistance. *J. Urban Econ.* 55 (2), 195–214.
- Early, D.W., Olsen, E.O., 2002. Subsidized housing, emergency shelters, and homelessness: an empirical investigation using data from the 1990 census. *Advances Econ. Anal. Policy* 2 (1), 1–34.
- Early, D.W., Olsen, E.O., 2012. Geographic price variation, housing assistance, and poverty. In: Jefferson, P.N. (Ed.), *Oxford Handbook of the Economics of Poverty*. Oxford University Press, Oxford.
- Ellen, I.G., O’Flaherty, B., 2007. Social programs and household size: evidence from New York City. *Popul. Res. Policy Rev.* 26 (4), 387–409.
- Ellen, I.G., Voicu, I., 2006. Nonprofit housing and neighborhood spillovers. *J. Policy Anal. Manage.* 25 (1), 31–52.
- Ellen, I.G., Schwartz, A.E., Voicu, I., Schill, M.H., 2007. Does federally subsidized rental housing depress neighborhood property values? *J. Policy Anal. Manage.* 26 (2), 257–280.
- Ellen, I.G., Tye, J.N., Willis, M.A., 2011. The secondary market for housing finance in the United States. In: Smith, M., Wachter, S.M. (Eds.), *The American Mortgage System Crisis and Reform*. University of Pennsylvania Press, Philadelphia, PA.
- Ellen, I.G., Lens, M.C., O’Regan, K., 2012. American murder mystery revisited: do housing voucher households cause crime? *Hous. Policy Debate* 22 (4), 551–572.
- Epple, D., Sieg, H., 1999. Estimating equilibrium models of local jurisdictions. *J. Polit. Econ.* 107 (4), 645–681.
- Eriksen, M.D., 2009. The market price of low-income housing tax credits. *J. Urban Econ.* 66 (2), 141–149. (in press).
- Eriksen, M.D., Rosenthal, S.S., 2010. Crowd out effects of place-based subsidized rental housing: new evidence from the LIHTC program. *J. Public Econ.* 94 (11–12), 953–966.
- Eriksen, M.D., Ross, A., 2013. The impact of housing vouchers on mobility and neighborhood attributes. *Real Estate Econ.* 41 (2), 255–277.
- Eriksen, M.D., Ross, A., Forthcoming. “Housing vouchers and the price of rental housing.” *Am. Econ. J. Econ. Policy*. (in press).
- Fertig, A.R., Reingold, D.A., 2007. Public housing, health, and health behaviors: is there a connection? *J. Policy Anal. Manage.* 26 (4), 831–859.
- Fetter, D.K., 2013a. How do mortgage subsidies affect home ownership? Evidence from the mid-century GI bills. *Am. Econ. J. Econ. Policy* 5 (2), 111–147.
- Fetter, D.K., 2013b. The Home Front: Rent Control and the Rapid Wartime Increase in Homeownership (unpublished manuscript).
- Finkel, M., DeMarco, D., Morse, D., Nolden, S., Rich, K., 1999. Status of HUD-Insured (Or Held) Multi-family Rental Housing in 1995: Final Report. Abt Associates Inc., Cambridge, MA.

- Fischel, W.A., 2000. Zoning and land use regulation. In: Bouckaert, B., De Geest, G. (Eds.), *Encyclopedia of Law and Economics*. In: *Civil Law and Economics*, vol. 2. Edward Elgar, Northampton, MA.
- Footo, C.L., Gerardi, K., Willen, P.S., 2008. Negative equity and foreclosure: theory and evidence. *J. Urban Econ.* 64 (2), 234–245.
- Footo, C.L., Gerardi, K.S., Willen, P.S., 2012. Why did so many people make bad decisions? The causes of the foreclosure crisis. In: Blinder, A.S., Lo, A.W., Solow, R.M. (Eds.), *Rethinking the Financial Crisis*. Russell Sage Foundation, New York, NY (Chapter 6).
- Footo, C., Gerardi, K., Willen, P., 2013. Government Policy and the Crisis: The Case of the Community Reinvestment Act. *Real Estate Research Blog*, Federal Reserve Bank of Atlanta (August 1).
- Freedman, M., Owens, E.G., 2011. Low-income housing development and crime. *J. Urban Econ.* 70 (2–3), 115–131.
- Friedman, J., Weinberg, D.H., 1982. *The Economics of Housing Vouchers*. Academic Press, New York, NY.
- Friedman, J., Weinberg, D.H., 1983. *The Great Housing Experiment*. Sage Publications, Beverly Hills, CA.
- Gabriel, S., 2001. Opening the doors to homeownership: challenges to federal policy. *Cityscape* 5 (2), 31–41.
- Gabriel, S.A., Rosenthal, S.S., 2009. Government-sponsored enterprises, the Community Reinvestment Act, and home ownership in targeted underserved neighborhoods. In: Glaeser, E.L., Quigley, J.M. (Eds.), *Housing Markets and the Economy: Risk, Regulation and Policy*. Lincoln Institute of Land Policy, Cambridge, MA.
- Gabriel, S., Rosenthal, S., 2010. Do the GSEs expand the supply of mortgage credit? New evidence of crowd out in the secondary mortgage market. *J. Public Econ.* 94, 953–966.
- Gabriel, S., Rosenthal, S., 2015. The boom, the bust and the future of homeownership. *Real Estate Econ.* (forthcoming).
- Galiani, S., Murphy, A., Pantano, J., 2012. *Estimating Neighborhood Choice Models: Lessons from a Housing Assistance Experiment*. Washington University (unpublished manuscript).
- Galster, G.C., Smith, R.E., Tatian, P.A., 1999a. *Assessing Property Value Impacts of Dispersed Housing Subsidy Programs: Final Report*. U.S. Department of Housing and Urban Development, Office of Policy Development and Research, Washington, DC.
- Galster, G.C., Tatian, P., Smith, R., 1999b. The impact of neighbors who use section 8 certificates on property values. *Hous. Policy Debate* 10 (4), 879–917.
- GAO, 2001. *Federal housing programs: what they cost and what they provide*: GAO-01-901R. GAO, Washington, DC.
- GAO, 2002. *Federal housing assistance: comparing the characteristics and costs of housing programs*: GAO-02-76. GAO, Washington, DC.
- Gerardi, K., Lambie-Hanson, L., Willen, P.S., 2013. Do borrower rights improve borrower outcomes? Evidence from the foreclosure process. *J. Urban Econ.* 73, 1–17.
- Ghent, A.C., Kudlyak, M., 2011. Recourse and residential mortgage default: evidence from U.S. States. *Rev. Financ. Stud.* 24 (9), 3139–3186.
- Ghent, A.C., Hernandez-Murillo, R., Owyang, M.T., 2013. *Did Affordable Housing Legislation Contribute to the Subprime Securities Boom?* (unpublished mimeo, August).
- Gibbons, S., Manning, A., 2006. The incidence of UK housing benefit: evidence from the 1990s reforms. *J. Public Econ.* 90 (4–5), 799–822.
- Glaeser, E.L., 2010. Housing policy in the wake of the crash. *Daedalus* 139 (4), 95–106.
- Glaeser, E.L., Gyourko, J., 2008. *Rethinking Federal Housing Policy*. AEI Press, Washington, DC.
- Glaeser, E.L., Luttmer, E.F.P., 2003. The misallocation of housing under rent control. *Am. Econ. Rev.* 93 (4), 1027–1046.
- Goodman, J.L., Nichols, J.B., 1997. Does FHA increase home ownership or just accelerate it? *J. Hous. Econ.* 6, 184–202.
- Grebler, L., Blank, D.M., Winnick, L., 1956. *Capital Formation in Residential Real Estate*. National Bureau of Economic Research, Princeton University Press, Princeton, NJ.
- Green, R.K., 2014. *Introduction to Mortgages and Mortgage Backed Securities*. Academic Press, Amsterdam.
- Green, R.K., Wachter, S.M., 2005. The American mortgage in historical and international context. *J. Econ. Perspect.* 19 (4), 93–114.

- Green, R.K., White, M.J., 1997. Measuring the benefits of homeownership: effects on children. *J. Urban Econ.* 41 (3), 441–461.
- Gustafson, J., Walker, J.C., 2002. Analysis of State Qualified Allocation Plans for the Low-Income Housing Tax Credit Program. Urban Institute, Washington, DC, U.S. Department of Housing and Urban Development.
- Gyourko, J., Molloy, R., 2015. Regulation. In: Duranton, G., Vernon Henderson, J., Strange, W. (Eds.), *Handbook of Regional and Urban Economics*, vol. 5. North-Holland, Amsterdam.
- Gyourko, J., Sinai, T., 2003. The spatial distribution of housing-related ordinary income tax benefits. *Real Estate Econ.* 31 (4), 527–575.
- Hammond, C.H., 1987. *The Benefits of Subsidized Housing Programs: An Intertemporal Approach*. Cambridge University Press, Cambridge, England.
- Hardaway, R., 2009. The great American housing bubble: re-examining cause and effect. *U. Dayton L. Rev.* 35 (1), 33–59.
- Harding, J.P., Rosenblatt, E., Yao, V.W., 2012. The foreclosure discount: myth or reality? *J. Urban Econ.* 74, 204–218.
- Harkness, J., Newman, S.J., 2003. Differential effects of homeownership on children from higher- and lower-income families. *J. Hous. Res.* 14 (1), 1–19.
- Harrison, D.M., Archer, W.R., Ling, D.C., Smith, M.T., 2002. Mitigating information externalities in mortgage markets: the role of Government-sponsored enterprises. *Cityscape* 6 (1), 115–143.
- Hartley, D., 2010. The effects of foreclosures on owner-occupied housing prices: supply or dis-amenity?: Federal Reserve Bank of Cleveland Working paper.
- Hembre, E., 2014. HAMP, Home Attachment, and Mortgage Default.
- Herbert, C.E., McCue, D.T., Sanchez-Moyano, R., 2014. Is homeownership still an effective means of building wealth for low-income and minority households? (Was it ever?). In: Belsky, E.S., Herbert, C.E., Molinsky, J.H. (Eds.), *Homeownership Built to Last: Balancing Access, Affordability, and Risk After the Housing Crisis*. Joint Center for Housing Studies, Harvard University, Cambridge, MA.
- Hilber, C.A.L., Mayer, C., 2009. Why do households without children support local public schools? Linking house price capitalization to school spending. *J. Urban Econ.* 65 (1), 74–90.
- Hilber, C.A.L., Turner, T.M., 2014. The mortgage interest deduction and its impact on homeownership decisions. *Rev. Econ. Stat.* 96 (4), 618–637.
- Hills, J., 1991. Distributional effects of housing subsidies in the United Kingdom. *J. Public Econ.* 44 (3), 321–352.
- Hilton, R., Hanson, C., Anderson, J., Finkel, M., Lam, K., Khadduri, J., Wood, M., 2004. Evaluation of the Mark-to-Market Program. U.S. Department of Housing and Urban Development, Office of Policy Development and Research, Washington, DC.
- Ho, G., Pennington-Cross, A., 2007. The varying effects of predatory lending laws on high-cost mortgage applications. *Fed. Reserve Bank St. Louis Rev.* 89 (1), 39–59.
- Horn, K.M., O'Regan, K.M., 2011. The low income housing tax credit and racial segregation. *Hous. Policy Debate* 21 (3), 443–473.
- Immergluck, D., 2009. *Foreclosed: High Risk Lending, Deregulation, and the Undermining of America's Mortgage Market*. Cornell University Press, Ithaca, NY.
- Jacob, B.A., 2004. Public housing, housing vouchers, and student achievement: evidence from public housing demolitions in Chicago. *Am. Econ. Rev.* 94 (1), 233–258.
- Jacob, B.A., Ludwig, J., 2012. The effects of housing assistance on labor supply: evidence from a voucher lottery. *Am. Econ. Rev.* 102 (1), 272–304.
- Jacob, B., Kapustin, M., Ludwig, J., 2014. Human capital effects of anti-poverty programs: evidence from a randomized housing voucher lottery: NBER Working paper 20164.
- Jaffee, D., Quigley, J.M., 2013. The future of the government sponsored enterprises: the role for government the U.S. mortgage market. In: Glaeser, E.L., Sinai, T. (Eds.), *Housing and the Financial Crisis*. University of Chicago Press.
- Kaufman, A., 2014. The influence of Fannie and Freddie on mortgage loan terms. *Real Estate Econ.* 42 (2), 472–496.

- Kennedy, S.D., 1980. Final Report of Housing Allowance Demand Experiment. Abt Associates Inc., Cambridge, MA.
- Kling, J.R., Liebman, J.B., Katz, L.F., 2007. Experimental analysis of neighborhood effects. *Econometrica* 75 (1), 83–119.
- Laferriere, A., Le Blanc, D., 2004. How do housing allowances affect rents? An empirical analysis of the french case. *J. Hous. Econ.* 13 (1), 36–67.
- Lang, B.J., 2012. Location incentives in the low-income housing tax credit: are qualified census tracts necessary? *J. Hous. Econ.* 21 (2), 142–150.
- Le Blanc, D., Laferriere, A., 2001. The effect of public social housing on households' consumption in France. *J. Hous. Econ.* 10 (4), 429–455.
- Lee, C.-M., Culhane, D.P., Watcher, S.M., 1999. The differential impacts of federally assisted housing programs on nearby property values: a Philadelphia case study. *Hous. Policy Debate* 10 (1), 75–93.
- Leger, M.L., Kennedy, S.D., 1990. Final Comprehensive Report of the Freestanding Housing Voucher Demonstration, vols. 1 & 2 Abt Associates Inc., Cambridge, MA.
- Lens, M.C., Ellen, I.G., O'Regan, K., 2011. Do vouchers help low-income households live in safer neighborhoods? Evidence on the housing choice voucher program. *Cityscape* 13 (3), 135–159.
- Liebowitz, S.J., 2009. Anatomy of a train wreck: causes of the mortgage meltdown. In: Holcomb, R., Powell, B. (Eds.), *Housing America: Building Out of a Crisis*. Transaction Publishers, New York.
- Lowry, Ira S. (Ed.), 1983. *Experimenting with Housing Allowances: The Final Report of the Housing Assistance Supply Experiment*. Oelgeschlager, Gunn & Hain, Cambridge, MA.
- Lui, H.-K., Suen, W., 2011. The effects of public housing on internal mobility in Hong Kong. *J. Hous. Econ.* 20 (1), 15–29.
- Malpezzi, S., Vandell, K., 2002. Does the low-income housing tax credit increase the supply of housing? *J. Hous. Econ.* 11 (4), 360–380.
- Mansur, E.T., Quigley, J.M., Raphael, S., Smolensky, E., 2002. Examining policies to reduce homelessness using a general equilibrium model of the housing market. *J. Urban Econ.* 52 (2), 316–340.
- Mayer, C., Pence, K., Sherlund, S.M., 2009. The rise in mortgage defaults. *J. Econ. Perspect.* 23 (1), 27–50.
- Mayo, S.K., Mansfield, S., Warner, D., Zwetchkenbaum, R., 1980. *Housing Allowances and Other Rental Assistance Programs—A Comparison Based on the Housing Allowance Demand Experiment, Part 2: Costs and Efficiency*. Abt Associates Inc., Cambridge, MA.
- McKenzie, J.A., 2002. A reconsideration of the jumbo/non-jumbo mortgage rate differential. *J. Real Estate Financ. Econ.* 25 (2/3), 197–213.
- Mian, A., Sufo, A., Trebbi, F., 2011. Foreclosures, house prices, and the real economy. NBER Working paper 16685.
- Mills, E.S., Sullivan, A., 1981. Market effects. In: Bradbury, K.L., Downs, A. (Eds.), *Do Housing Allowances Work?* Brookings Institution, Washington, DC.
- Morgan, D.P., Iverson, B., Botsch, M., 2012. Subprime foreclosures and the 2005 bankruptcy reform. *Fed. Reserve Bank New York Econ. Pol. Rev.* 18 (1), 47–57.
- Moulton, S., 2014. Did affordable housing mandates cause the subprime mortgage crisis? *J. Hous. Econ.* 24, 21–38.
- Mulligan, C.B., 2010. Foreclosures, enforcement, and collections under the federal mortgage modification guidelines: NBER Working paper 15777.
- Munnell, A.H., Tootell, G.M.B., Browne, L.E., McEneaney, J., 1996. Mortgage lending in Boston: interpreting HMDA data. *Am. Econ. Rev.* 86 (1), 25–53.
- Murray, M.P., 1975. The distribution of tenant benefits in public housing. *Econometrica* 43 (4), 771–788.
- Murray, M.P., 1983. Subsidized and unsubsidized housing starts: 1961–1977. *Rev. Econ. Stat.* 65 (4), 590–597.
- Murray, M.P., 1999. Subsidized and unsubsidized housing stocks 1935 to 1987: crowding out and cointegration. *J. Real Estate Financ. Econ.* 18 (1), 107–124.
- National Research Council, Committee to Evaluate the Research Plan of the Department of Housing and Urban Development, Center for Economic, Governance, and International Studies, Division of

- Behavioral and Social Sciences and Education, 2008. Rebuilding the Research Capacity at HUD. The National Academies Press, Washington, DC.
- Newman, S., Harkness, J., 2000. Assisted housing and the educational attainment of children. *J. Hous. Econ.* 9 (1–2), 40–63.
- Newman, S.J., Harkness, J.M., 2002. The long-term effects of public housing on self-sufficiency. *J. Policy Anal. Manage.* 21 (1), 21–43.
- Newman, S.J., Schnare, A.B., 1997. “... and a Suitable Living Environment”: the failure of housing programs to deliver on neighborhood quality. *Hous. Policy Debate* 8 (4), 703–741.
- Newman, S., Holupka, C.S., Harkness, J., 2009. The long-term effects of housing assistance on work and welfare. *J. Policy Anal. Manage.* 28 (1), 81–101.
- Nichols, M.W., Hendrickson, J.M., Griffith, K., 2011. Was the financial crisis the result of ineffective policy and too much regulation? An empirical investigation. *J. Bank. Regul.* 12, 236–251.
- Office of Policy Development and Research of the U.S. Department of Housing and Urban Development. 2010. “Report to Congress on the Root Causes of the Foreclosure Crisis”.
- Olsen, E.O., 1998. Introduction to economics of rent control. *Reg. Sci. Urban Econ.* 28 (6), 673–678.
- Olsen, E.O., 2003. Housing programs for low-income households. In: Moffitt, R.A. (Ed.), *Means-Tested Transfer Programs in the United States*. University of Chicago Press, Chicago.
- Olsen, E.O., 2007a. Promoting homeownership among low-income households: Opportunity and Ownership Project, Report Number 2. Urban Institute, Washington, DC.
- Olsen, E.O., 2007b. A primer on U.S. housing markets and housing policies: a review article. *Reg. Sci. Urban Econ.* 37 (5), 618–624.
- Olsen, E.O., 2008. Getting more from low-income housing assistance: The Hamilton Project, Discussion Paper 2008-13. The Brookings Institution, Washington, DC. http://www.brookings.edu/papers/2008/09_low_income_housing_olsen.aspx.
- Olsen, E.O., 2009. The cost-effectiveness of alternative methods of delivering housing subsidies. University of Virginia, Charlottesville, VA. <http://economics.virginia.edu/sites/economics.virginia.edu/files/CESurvey2009.pdf> (unpublished manuscript).
- Olsen, E.O., Ludwig, J., 2013. The performance and legacy of housing policies. In: Bailey, M., Danziger, S. (Eds.), *The Legacies of the War on Poverty*. Russell Sage Foundation, New York.
- Olsen, E.O., Tyler, C.A., King, J.W., Carrillo, P.E., 2005. The effects of different types of housing assistance on earnings and employment. *Citiescape* 8 (2), 163–187.
- ORC/Macro, 2001. Quality Control for Rental Assistance Subsidies Determinations. U.S. Department of Housing and Urban Development, Office of Policy Development and Research, Washington, DC.
- Oreopoulos, P., 2003. The long-run consequences of living in a poor neighborhood. *Q. J. Econ.* 118 (4), 1533–1575.
- Orr, L., Feins, J.D., Jacob, R., Beecroft, E., Sanbonmatsu, L., Katz, L., Liebman, J.B., Kling, J.R., 2003. Moving to Opportunity for Fair Housing Demonstration Program: Interim Impacts Evaluation. US Department of Housing and Urban Development, Washington, DC.
- Painter, G., Redfern, C., 2002. The role of interest rates in influencing long-run homeownership rates. *J. Real Estate Financ. Econ.* 25 (2/3), 243–267.
- Passmore, W., Sherlund, S.M., Burgess, G., 2005. The effect of housing government-sponsored enterprises on mortgage rates. *Real Estate Econ.* 33, 427–463.
- Pence, K.M., 2006. Foreclosing on opportunity: state laws and mortgage credit. *Rev. Econ. Stat.* 88 (1), 177–182.
- Pennington-Cross, A., 2010. The duration of foreclosures in the subprime mortgage market: a competing risks model with mixing. *J. Real Estate Financ. Econ.* 40 (2), 109–129.
- Popkin, S.J., Rich, M.J., Hendey, L., Hayes, C., Parilla, J., Galster, G., 2012. Public Housing transformation and Crime: making the Case for responsible relocation. *Citiescape* 14 (3), 137–160.
- Poterba, J.M., Sinai, T., 2011. Revenue costs and incentive effects of the mortgage interest deduction for owner-occupied housing. *Natl. Tax J.* 64 (2), 531–564.
- Quercia, R., McCarthy, G., Wachter, S., 2003. The impacts of affordable lending efforts on homeownership rates. *J. Hous. Econ.* 12 (1), 29–59.

- Reeder, W.J., 1985. The benefits and costs of the section 8 existing housing program. *J. Public Econ.* 26 (3), 349–377.
- Reid, C., Laderman, E., 2011. Constructive credit: revisiting the performance of community reinvestment act lending during the subprime crisis. In: Wachter, S.M., Smith, M.M. (Eds.), *The American Mortgage System: Crisis and Reform*. University of Pennsylvania Press, Philadelphia, PA.
- Reid, C., Willis, M.A., Seidman, E., Ding, L., Silver, J., Ratchliffe, J., 2013. Debunking the CRA Myth—again: Research Report. University of North Carolina Center (UNC) for Community Capital.
- Rose, J.D., 2011. The incredible HOLC? Mortgage relief during the Great Depression. *J. Money, Credit, Bank.* 43 (6), 1073–1107.
- Rose, J.D., Snowden, K.A., 2013. The New Deal and the origins of the modern American real estate loan contract. *Explor. Econ. Hist.* 50 (4), 548–566.
- Rosen, H.S., 1979. Housing decisions and the U.S. income tax: an econometric analysis. *J. Public Econ.* 11 (1), 1–23.
- Rosen, H.S., 1985. Housing subsidies: effects on housing decisions, efficiency, and equity. In: Auerbach, A. J., Feldstein, M. (Eds.), *Handbook of Public Economics*, vol. 1. North-Holland, Amsterdam.
- Rosenthal, S.S., 2014. Are private markets and filtering a viable source of low-income housing? Estimates from a ‘repeat income’ model. *Am. Econ. Rev.* 104 (2), 687–706.
- Rossi-Hansberg, E., Sarte, P.-D., Raymond Owens, I.I.I., 2010. Housing externalities. *J. Polit. Econ.* 118 (3), 485–535.
- Rydell, C.P., Neels, K., Lance Barnett, C., 1982. Price Effects of a Housing Allowance Program. The Rand Corporation, Santa Monica, CA.
- Sanbonmatsu, L., Ludwig, J., Katz, L.F., Gennetian, L.A., Duncan, G.J., Kessler, R.C., Adam, E., McDade, T.W., Lindau, S.T., 2011. Moving to Opportunity for Fair Housing Demonstration Program: Final Impacts Evaluation. U.S. Department of Housing and Urban Development, Office of Policy Development and Research, Washington, DC.
- Santiago, A.M., Galster, G.C., Tatian, P., 2001. Assessing the property value impacts of the dispersed housing subsidy program in Denver. *J. Policy Anal. Manage.* 20 (1), 65–88.
- Schill, M.H., 2005. Regulations and housing development: what we know. *Cityscape* 8 (1), 5–19.
- Schone, B.S., 1992. Do means tested transfers reduce labor supply? *Econ. Lett.* 40 (3), 353–357.
- Schuetz, J., Been, V., Ellen, I.G., 2008. Neighborhood effects of concentrated mortgage foreclosures. *J. Hous. Econ.* 17 (4), 306–319.
- Schwartz, A.E., Ellen, I.G., Voicu, I., Schill, M.H., 2006. The external effects of place-based subsidized housing. *Reg. Sci. Urban Econ.* 36 (6), 679–707.
- Sherlund, S., 2008. The Jumbo-Conforming Spread: A Semi-parametric Approach (unpublished manuscript).
- Shlay, A.B., 2006. Low-income homeownership: American dream or delusion? *Urban Stud.* 43 (3), 511–531.
- Shroder, M., 2002a. Locational constraint, housing counseling, and successful lease-up in a randomized housing voucher experiment. *J. Urban Econ.* 51 (2), 315–338.
- Shroder, M., 2002b. Does housing assistance perversely affect self-sufficiency? A review essay. *J. Hous. Econ.* 11 (4), 381–417.
- Shroder, M., Reiger, A., 2000. Vouchers versus production revisited. *J. Hous. Res.* 11 (1), 91–107.
- Sinai, T., Waldfoegel, J., 2005. Do low-income housing subsidies increase the occupied housing stock? *J. Public Econ.* 89 (11–12), 2137–2164.
- Snowden, K.A., 2006. Mortgage foreclosures and delinquencies: 1926–1979. In: Carter, S.B., Sigmund Gartner, S., Haines, M.R., Olmstead, A.L., Sutch, R., Wright, G. (Eds.), *Historical Statistics of the United States, Earliest Times to the Present: Millennial Edition*. Cambridge University Press, New York, Table Dc1255–1270.
- Struyk, R.J., Bendick, M. (Eds.), 1981. *Housing Vouchers for the Poor: Lessons from a National Experiment*. The Urban Institute Press, Washington, DC.
- Susin, S., 2002. Rent vouchers and the price of low-income housing. *J. Public Econ.* 83 (1), 109–152.
- Susin, S., 2005. Longitudinal outcomes of subsidized housing recipients in matched survey and administrative data. *Cityscape* 8 (2), 189–218.

- Szymanoski, E., Reeder, W., Raman, P., Comeau, J., 2012. The FHA single-family insurance program: performing a needed role in the housing finance market: Housing Finance Working Paper Series. U.S. Department of Housing and Urban Development/Office of Policy Development and Research.
- Turnham, J., Herbert, C., Nolden, S., Feins, J., Bonjorni, J., 2004. Study of homebuyer activity through the HOME Investment Partnership program: Abt Associates Report. U.S. Department of Housing and Urban Development, Office of Policy Development and Research, Washington, DC.
- U.S. Department of Housing and Urban Development, 1980. Experimental Housing Allowance Program: The 1980 Report. Government Printing Office, Washington, DC.
- U.S. General Accounting Office, 1997. Tax credits: opportunities to improve oversight of the low-income housing program: GGD/RCED-97-55. GAO, Washington, DC.
- Van Order, R., Yezer, A.M., 2014. FHA: recent history and future prospects. *Hous. Policy Debate* 24 (3), 644–650.
- Vandell, K.D., 1995. FHA restructuring proposals: alternatives and implications. *Hous. Policy Debate* 6, 299–394.
- Vigdor, J.L., 2006. Liquidity constraints and housing prices: theory and evidence from the VA mortgage program. *J. Public Econ.* 90 (8–9), 1579–1600.
- Voicu, I., Paley, I., Lopez, A.E., Fang, I., 2014. Information externalities, neighborhood characteristics, and home mortgage pricing and underwriting. *Real Estate Econ.* (forthcoming).
- von Hoffman, A., 2012. A rambling edifice: American housing policy in the twentieth century: Working paper W12-9. Harvard University, Joint Center for Housing Studies.
- Wallace, J.E., Bloom, S.P., Holshouser, W.L., Mansfield, S., 1981. Participation and Benefits in the Urban Section 8 Program: New Construction and Existing Housing, vols. 1 & 2. Abt Associates, Inc., Cambridge, MA.
- Wallison, P.J., 2009. “Dissenting Statement,” The Financial Crisis Inquiry Report: Final Report of the National Commission on the Causes of the Financial and Economic Crisis in the United States.
- Walters, C.R., 2009. Do subsidized housing units depreciate faster than unsubsidized ones? *J. Hous. Econ.* 18 (1), 49–58.
- Weicher, J.C., 2012. Housing Policy at a Crossroads: The Why, How, and Who of Assistance Programs. AEI Press, Washington, DC.
- Whitehead, C., Scanlon, K. (Eds.), 2007. Social Housing in Europe. London School of Economics and Political Science, London.
- Wilner, D.M., Walkley, R.P., Pinkerton, T.C., Tayback, M., 1962. The Housing Environment and Family Life: A Longitudinal Study of the Effects of Housing on Morbidity and Mental Health. Johns Hopkins Press, Baltimore.
- Wong, Y.-C., Liu, P.-W., 1988. The distribution of benefits among public housing tenants in Hong Kong and related policy issues. *J. Urban Econ.* 23 (1), 1–20.
- Yinger, J., 1995. Closed Doors, Opportunities Lost: The Continuing Costs of Housing Discrimination. Russell Sage Foundation, New York, NY.

CHAPTER 15

How Mortgage Finance Affects the Urban Landscape

Sewin Chan*, Andrew Haughwout†, Joseph Tracy†

*Robert F. Wagner School of Public Service, New York University, NY, USA

†Federal Reserve Bank of New York, NY, USA

Contents

15.1. Mortgage Finance in the United States	989
15.1.1 Important design features of mortgage finance	991
15.1.1.1 <i>Collateral</i>	991
15.1.1.2 <i>Lien priority</i>	993
15.1.1.3 <i>Recourse</i>	994
15.1.1.4 <i>Assumability and portability</i>	994
15.1.1.5 <i>Mortgage term</i>	994
15.1.2 Withdrawing home equity	995
15.2. How Mortgage Finance Affects the Market for Owner-Occupied Housing	997
15.2.1 The demand for owner-occupied housing	997
15.2.2 The effect of demand shifts on price and quantity	1000
15.3. The Distribution of Mortgage Credit	1005
15.3.1 Access to mortgage credit	1006
15.3.2 The distribution of high-cost lending	1008
15.3.3 Mortgage finance over the life cycle	1015
15.3.3.1 <i>Transitions to homeownership</i>	1016
15.3.3.2 <i>Mortgage finance at older ages</i>	1018
15.4. Negative Equity	1022
15.4.1 Negative equity and strategic default	1023
15.4.2 Negative equity and housing turnover	1024
15.4.3 Negative equity and housing investment	1030
15.5. Foreclosures	1034
15.5.1 The distribution of foreclosures	1034
15.5.2 Foreclosure externalities	1037
15.6. Conclusion	1039
Acknowledgments	1040
References	1040

Abstract

This chapter considers the structure of mortgage finance in the United States and its role in shaping patterns of homeownership, the nature of the housing stock, and the organization of residential activity. We start by providing some background on the design features of mortgage contracts that distinguish them from other loans and that have important implications for issues presented in the

rest of the chapter. We then explain how mortgage finance interacts with public policy, particularly tax policy, to influence a household's decision to own or rent and how shifts in the demand for owner-occupied housing are translated into housing prices and quantities, given the unusual nature of housing supply. We consider the distribution of mortgage credit in terms of access and price, by race, ethnicity, and income, and over the life cycle, with particular attention to the role of recent innovations such as nonprime mortgage securitization and reverse mortgages. The extent of negative equity has been unprecedented in the past decade, and we discuss its impact on strategic default, housing turnover, and housing investment. We describe spatial patterns in foreclosure and summarize the evidence for foreclosure spillovers in urban neighborhoods. Finally, we offer some thoughts on future innovations in mortgage finance.

Keywords

Mortgage finance, Mortgage contract, Housing market, Homeownership, Foreclosure, Default, Negative equity, Mobility, Housing maintenance, Housing investment, Neighborhoods, Metropolitan areas

JEL Classification Codes

D1: Household Behavior and Family Economics; G21: Banks; Other Depository Institutions; Micro Finance Institutions; Mortgages; Foreclosures; R1: General Regional Economics, R2: Household Analysis

Over 80% of Americans live in metropolitan areas, and housing is the dominant land use in cities. For many Americans, homeownership is an important goal, and a substantial majority of the population, including renters, believes that homeownership is a good way to improve their financial situation (Fannie Mae, 2013). For many owners, and for the great majority of renters, purchasing a home will mean obtaining a mortgage. For this reason, the availability and form of mortgages is an important determinant of the homeownership rate, which in turn affects the nature of the housing stock and the organization of residential activity within and across metropolitan areas.

In this chapter, we consider the literature on mortgage finance in the United States and its role in shaping the urban landscape. The 2000s witnessed an enormous boom/bust cycle in the residential real estate market, followed by the sharpest contraction in the overall economy since the 1930s. These events, which are widely thought to have been driven at least in part by the mortgage market, had a pronounced spatial pattern that research is only beginning to completely understand. Our workhorse models of local demand and supply of owner-occupied housing can give us only partially satisfactory explanations for the patterns we observe in the data, and more work, both theoretical and empirical, is needed to understand why the boom/bust cycle occurred when and where it did. For example, the user cost framework that has long served as the basis for analyzing how credit conditions affect the demand for owner-occupied housing provides a good basis for understanding the *direction* of demand shifts during the boom, but the basic model requires significant extensions to capture the *magnitude* and *locational patterns* of the demand volatility we observed during the 2000s.

Similarly, the decade produced some challenges for our understanding of how demand shifts are translated into prices and quantities. Why did prices rise so much in housing markets that by most measures featured relatively elastic supply? Expectations and buyers with motivations other than owner-occupancy are beginning to play important roles in understating these dynamics. Another key set of questions concern the spatial distribution of credit and its effects on the spatial dimensions of the boom and bust *within* metro areas. The long tradition of studies on the basic question of access to credit is undergoing a revival, this time with a special emphasis on nonprime credit. There is still much to learn on this topic, although our understanding in this area is constrained by inadequate data.

Finally, the enormous wave of foreclosures that followed the bust in housing prices featured spatial patterns that were perhaps even stronger than those found for lending in the boom. We are only now beginning to understand the complex dynamics of neighborhood housing prices under conditions of market distress, and work is likely to continue on this front for many years.

Our discussion in [Section 15.1](#) begins with a description of the evolution of the mortgage contract and its salient features. In [Section 15.2](#), we provide an analysis of the way that mortgage finance interacts with public policy—particularly the tax code—to influence the rent-own choice. How shifts in the relative demand for owner-occupancy are translated into housing prices and quantities requires understanding the unusual nature of housing supply, a topic that we also discuss in [Section 15.2](#). [Section 15.3](#) discusses the distribution of mortgage credit in metropolitan areas, across both geographies and demographics. Important innovations in recent years—nonprime mortgage securitization and reverse mortgages—have substantially influenced access to credit within metropolitan areas and have therefore had significant real impacts as well. An important feature of mortgages as they currently exist in the United States is that they do not have margin calls. For this reason, a decline in house prices can put borrowers in negative equity, a situation that has many potential consequences that we discuss in [Section 15.4](#). One of the most significant implications of house price declines is foreclosure, a phenomenon that has become enormously important in many metropolitan areas in the last decade. [Section 15.5](#) discusses reasons for spatial patterns in foreclosures and summarizes evidence for the “spillovers” associated with foreclosures in urban neighborhoods. [Section 15.6](#) provides a brief summary of potential reforms to mortgage contracts and points to future research directions.

15.1. MORTGAGE FINANCE IN THE UNITED STATES

Most homeowners finance their purchases with a mortgage. For the great majority of homeownership families, the house is a very significant part of their wealth, and the

availability of mortgage credit allows households to smooth their consumption rather than save for many years to buy a home with cash. For example, a household with the 2012 median income of \$53,046 would need to save nearly three and one-half years' worth of income to purchase the median-valued owner-occupied housing unit, which was valued at \$181,400.¹ At a saving rate of 20% per year, this household would save about \$10,600 per year, which (assuming an annual return on the savings of 5%) would allow the household to purchase the home after 13 years (assuming no change in income or house prices); savings at the national average rate of 4.5% of income would require 33 years to develop a fund large enough to purchase the house.

Mortgage credit allows the household to purchase its home with far less of an up-front consumption sacrifice. Assuming mortgage interest rates of 5% per annum, the median income household could qualify for a mortgage to buy the median house after saving a 20% down payment of \$36,280. This would take a little over 3 years with a saving rate of 20% of income each year. The household will be paying back the loan for the next 30 years but will be doing so while enjoying the benefits of homeownership rather than looking forward to them.

Figure 15.1 displays the homeownership rate over time and Figure 15.2 shows the prevalence of mortgage finance in recent years. The significant (albeit apparently temporary) rise in the homeownership rate during the housing boom of the late 1990s and early 2000s was associated with a very substantial increase in mortgage credit, underscoring the important relationship between the two series. In this section, we describe important

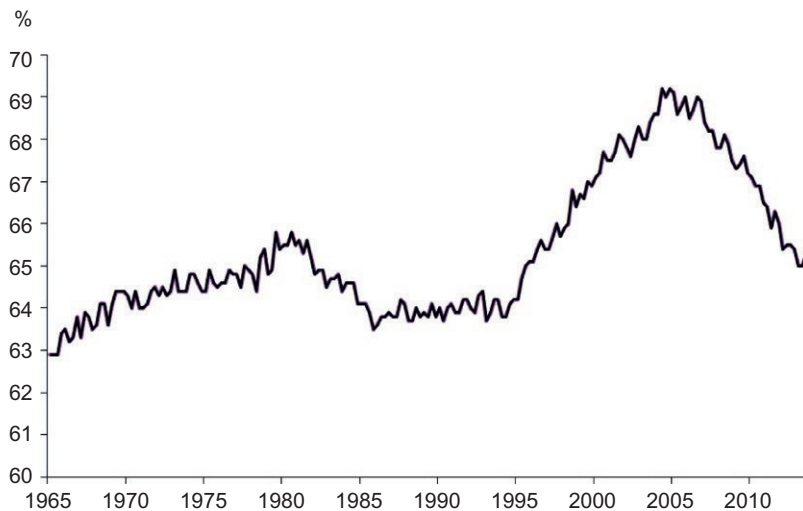


Figure 15.1 Homeownership rate. Source: US Census Bureau, *Current Population/Housing Vacancy Survey*, Series H-111. Fraction of housing units that are owner-occupied.

¹ Median values from the US Census, 2008–2012 averages.

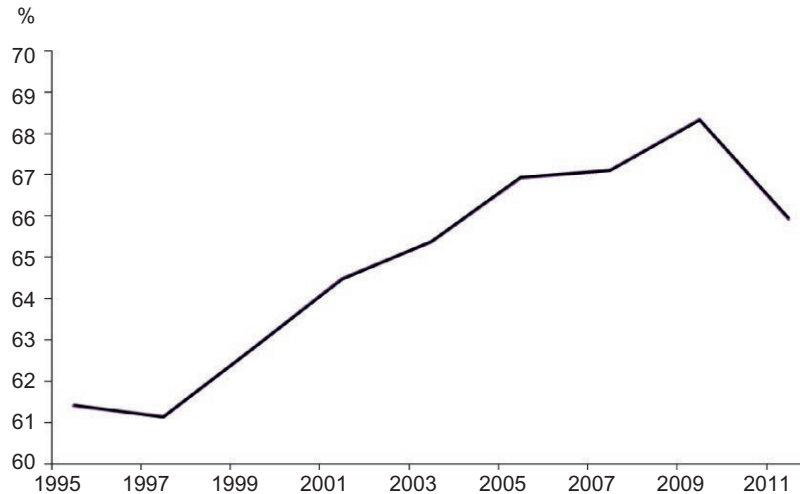


Figure 15.2 Fraction of homeowners with a mortgage. *Source: US Census Bureau, Current Housing Reports, Series H150/01, American Housing Survey. Fraction of owner-occupied housing units with a mortgage.*

features of mortgages in the United States and discuss how they affect important outcomes in the housing market and the risks they pose.

15.1.1 Important design features of mortgage finance

For historic and institutional reasons, mortgage finance in the United States has features that distinguish them from other loans. The mortgage market has long featured a prominent role for both explicit government influence and implicit government influence, from support for long-term, fixed-rate mortgage (FRM) contracts to limits on the ability of lenders to pursue deficiency judgments against borrowers who default on the mortgage contract. All of these features have effects on the costs and benefits of borrowing and lending in the mortgage market, and we outline the most salient of them below.

15.1.1.1 Collateral

Collateralized loans stipulate an interest rate or schedule of rates and margin requirements. The margin requirements consist of an initial margin and possibly a maintenance margin. The initial margin determines the degree to which the loan is overcollateralized at the outset. This provides protection to the lender in the event of a decline in the value of the collateral. A maintenance margin requirement specifies a minimum margin that the borrower must maintain over time. If the borrower fails to make a margin call, then the lender can call the loan. The combination of initial and maintenance margin requirements can significantly limit the possibility that the loan becomes “underwater” where the value of the collateral is less than the outstanding loan balance.

Mortgages have the feature that they only specify an initial margin—the down payment percentage—which determines the initial loan-to-value (LTV) ratio. Since there is no maintenance margin requirement on mortgages, if house prices decline subsequent to the mortgage origination, the borrower's equity is reduced and can become negative. As a consequence, negative equity is a more important consideration for mortgages than for other types of collateralized loans.

An alternative to a maintenance margin is to have mortgages with short terms. For example, the typical Canadian mortgage has a term of 5 years.² The borrower must roll over the mortgage at expiration. At the rollover date, a borrower must reestablish at least the minimum required level of equity. A downside to maintenance margins, or equivalently to short mortgage terms, is that the borrower may not be able to post the additional margin (and therefore cannot roll over the loan) and is forced into a sale or a default. This can lead to “fire sales” and place further downward pressure on prices.³ Lea (2010) provided data showing that short- or medium-term mortgages of the Canadian type are common in OECD countries, relative to longer term FRMs. The United States is distinguished by having a very large market share for long-term, FRMs. This will have implications for how the housing and mortgage markets react in times of stress.

The absence of a maintenance margin can create an asymmetry between the current owner and a potential new owner with regard to equity. If house prices decline pushing the current LTV below the minimum requirement at origination, the current owner by simply making the required monthly mortgage payments can continue to live in the house without adding additional equity. However, a potential new owner would have to satisfy the minimum LTV requirement. This means that a new owner would have to put more equity into the house than the existing owner currently has in the house. In this situation, the current owner would also have to put in additional cash in order to move to an equivalently priced home (abstracting from selling costs). This creates an incentive for the current owner to set a higher list price for the house even if this means taking a longer time to sell. Genesove and Mayer (1997) found evidence for this effect in the Boston condo market in the early 1990s.

For long-term mortgages, lenders should set the initial margin based on the perceived degree of house price risk. To maintain a relatively constant risk of negative equity, mortgage lenders would need to require a larger down payment percentage during periods when, or in locations where, house prices have a higher volatility. Evaluating whether this occurs in practice is difficult because we do not have data on the time series of beliefs

² One of the factors that impacts the short term is that a high fraction of mortgages are funded through deposits and that deposit guarantees in Canada cover terms up to 5 years. See Kiff (2009).

³ This rollover risk from short-term mortgages was a significant problem in the United States during the Depression and led to the development of long-term amortizing mortgages offered by the Federal Housing Administration (FHA). See Green and Wachter (2005).

Table 15.1 Combined LTV ratio at origination

Percentile	All housing purchases ^a				Nonprime purchase mortgages ^b			
	25th	50th	75th	90th	25th	50th	75th	90th
2004	56	80	95	100	80	95	100	100
2005	64	86	99	100	80	95	100	100
2006	70	90	100	100	90	99	100	100

^aGlaeser et al. (2013). DataQuick data from 89 metro areas.

^bLoan Performance data on securitized nonprime mortgages.

by mortgage lenders regarding house price risk.⁴ However, as shown in Table 15.1, the distribution of initial LTVs rose during the height of the most recent housing boom from 2004 to 2006.

15.1.1.2 Lien priority

A borrower may take out more than one loan that is collateralized by the same house. In this case, the lien priority is important in determining the order in which creditors have a claim on the proceeds from a sale of the house. The lien priority typically is based on the date in which the lien was recorded with the county, rather than the date that the loan was originated. That is, the earliest recorded existing lien has the top priority.⁵ This voluntary recording system provides a strong incentive for creditors to pay the recording fee in a prompt manner. However, the lien priority can create a potential friction for interventions aimed at assisting distressed borrowers. For example, consider a program that facilitates negative equity borrowers to refinance their mortgage and thus reduce their monthly payments.⁶ For borrowers with multiple liens, the lender with the first lien that is participating in this program would not want to do the refinance if this required giving up their first-lien position. Typically, the second-lien holders would have to agree to resubordinate their lien behind the refinanced mortgage in this case.⁷ But if the borrower's combined LTV across the first and second liens was low enough, then this friction disappears. In this case, the borrower could refinance by paying off all loans and credit lines secured by the house and replace them with a single refinanced mortgage.

⁴ For an interesting discussion on the heterogeneity of beliefs regarding the housing market during the boom, see Gerardi et al. (2010).

⁵ Exceptions to this rule are tax liens for unpaid property taxes that have superpriority regardless of the filing date.

⁶ Absent a special program, negative equity borrowers face a financial friction to refinancing in that they must have the financial resources to both fully pay off the existing mortgage and put a down payment on the refinance mortgage. See Caplin et al. (1997b) for estimates of the degree to which declining house prices can generate a friction to refinancing.

⁷ States can allow for "equitable subrogation" where the first-lien holder does not need an agreement from second-lien holders to resubordinate so long as a refinance of the first lien does not disadvantage the junior lien holders. Bond et al. (2013) found that states with equitable subrogation experienced a higher refinance rate holding other factors constant.

15.1.1.3 *Recourse*

Mortgage loans can be either nonrecourse or recourse. Where mortgages are nonrecourse, the lender only has claim to the sale proceeds from the house to satisfy the balance due on the mortgage. In contrast, where mortgage loans are recourse, if the proceeds from the sale do not fully pay off the balance on the mortgage, then the lender can file for a deficiency judgment against the borrower. The deficiency judgment allows the lender to use the borrower's other financial assets to pay off the remaining balance on the mortgage. Recourse varies by state, and for a few states, by the type of mortgage. This variation in the degree to which lenders are able to collect deficiencies appears to be unique to the United States. In [Lea's \(2010\)](#) study of 12 developed countries, in all of the other countries, mortgages are recourse and lenders typically pursue borrowers for deficiencies. Recourse may affect a borrower's behavior prior to a default as well as during a foreclosure.⁸

15.1.1.4 *Assumability and portability*

Typically, mortgage loans in the United States are neither assumable nor portable.⁹ A mortgage is assumable if the borrower can transfer the mortgage to the buyer upon the sale of the property, but the buyer must still meet the underwriting guidelines of the lender. If a mortgage is portable, then a borrower who sells and buys a property can transfer the existing mortgage to the new property, so long as the new property provides adequate collateralization for the mortgage. That is, upon the sale of the property with an assumable mortgage, the lender may end up retaining the collateral and re-underwriting a new borrower. In contrast, with a portable mortgage, the lender may end up retaining the borrower and re-underwriting the collateral. As we will discuss, the fact that conventional mortgages typically are neither assumable nor portable becomes important for FRMs in a rising interest rate environment.

15.1.1.5 *Mortgage term*

Another feature of most US mortgages is that they have long terms. This is important given that the housing collateral backing up the mortgage is a depreciating asset, where the rate of depreciation depends on actions taken by the borrower. An important offset to the depreciation rate is the extent to which the borrower makes costly investments in maintenance and improvements (hereafter housing investments) to the house. These expenditures typically are significant as measured relative to the value of the property.

⁸ [Ghent and Kudlyak \(2011\)](#) found that on average, US borrowers are 30% more likely to default in non-recourse states. [Jones \(1993\)](#) found for a sample of Canadian mortgages originated in Alberta and British Columbia between 1982 and 1986 that recourse reduces the incidence of strategic default.

⁹ An exception is FHA mortgages, which are assumable. In Canada, mortgages are portable, and in Alberta and British Columbia, they are also assumable ([Tractlet, 2010](#); [Crawford et al., 2013](#)). However, their short terms limit the value of this option. Mortgages in Denmark are also assumable ([Green and Wachter, 2005](#)).

Gyourko and Tracy (2006) using American Housing Survey (AHS) data from 1985 to 1993 reported an average expenditure level of \$2889 (in 2012 dollars), or 1.7% of the reported house value. Haughwout et al. (2013) using Consumer Expenditure Survey (CEX) data from 2007 to 2012 reported a mean expenditure of \$3152 (in 2012 dollars), or 1.4% of the house value. In aggregate, over the past 20 years, these housing investments averaged 45% of the value of new home construction. Maintaining incentives over time for the borrower to continue to make these costly investments is important for sustaining the value of the collateral.

15.1.2 Withdrawing home equity

Over time, a lender is typically protected by the buildup of equity through amortization of the loan balance and house price appreciation. However, in most cases, the lender cannot prevent the borrower from withdrawing this additional equity. This can expose the lender to risks even though the lender retains lien priority.

There are three separate ways that owners may borrow against their home equity to convert it into cash available for consumption or other purposes (Greenspan and Kennedy, 2008). The first is by refinancing their first-lien mortgage through a “cash-out” refinance. Here, a homeowner pays off their previous mortgage with the proceeds of a new, larger, mortgage. The difference between the balances on the two mortgages is available to the borrower as cash at settlement, and the borrower may use those proceeds as she likes, without restriction. Note that refinancing in this way will typically result in a new set of scheduled payments that are, in the early years of the contract, predominately interest, which as we will discuss is tax deductible.

The other methods of using home equity as collateral for borrowing involve either junior liens or a reverse mortgage, which we describe later. There are two forms of junior liens: closed-end seconds (CESs) and home equity lines of credits (HELOCs). A CES is a junior lien installment loan that has the same form as first mortgages. They deliver a fixed amount of up-front cash to the borrower and have a fixed term over which principal and interest are repaid to the lender. In contrast, HELOCs are revolving instruments that are more like credit cards. HELOCs typically have a borrowing limit, related to the total available housing equity, which the homeowner can utilize as she wishes during the “draw” period, typically 10 years.¹⁰

Junior liens, especially HELOCs, were an important part of the run-up in debt secured by housing during the boom. Aggregate HELOC balances are considerably larger than CESs and are typically held by borrowers with higher credit scores (Lee et al., 2013). Figure 15.3 shows the evolution of mortgages (including both first liens

¹⁰ At “end-of-draw,” no further borrowing is allowed. At that point, HELOCs begin to amortize and become more similar to variable rate closed-end mortgages. As of this writing, many HELOCs originated during the boom years are beginning to enter the end-of-draw period.

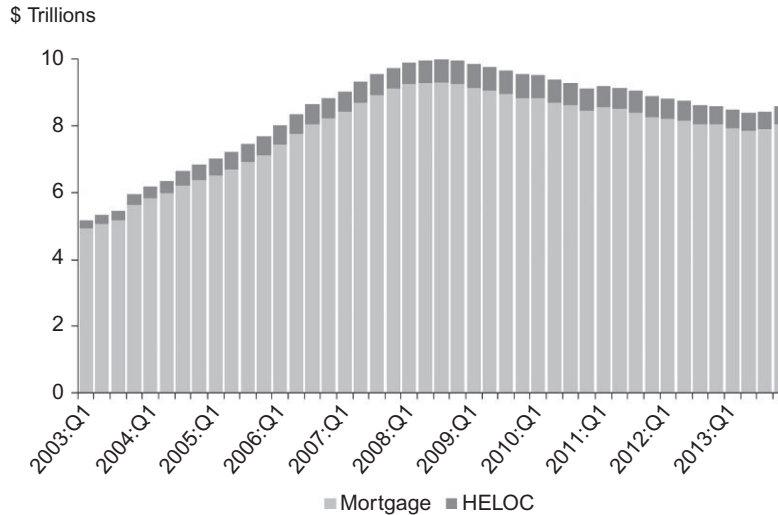


Figure 15.3 Mortgages and HELOCs. *Source: FRBNY Consumer Credit Panel.*

and CESs) and HELOC balances between 2003 and 2013. HELOC balances rose almost threefold from \$242 billion at the start of 2003 to over \$714 billion in 2009. Since 2009, HELOC balances have steadily declined and they stood at \$529 billion at the end of 2013.

Borrowers may put the proceeds from junior liens to many purposes. Junior liens may be used to increase leverage at the purchase of a new home (Greenspan and Kennedy, 2008; Lee et al., 2013). At the peak of the housing boom in 2006, nearly half of all home purchases in locations like Las Vegas and Miami involved second liens. The so-called subsequent seconds—junior liens added after the home purchase when the borrower has built some equity through the combination of amortization and rising home values—can be used for virtually any purpose, from remodeling the home, financing a college education, paying down unsecured credit balances, to taking a vacation (Mian and Sufi, 2011).

One important potential use of CESs and HELOCs is to allow equity extraction for the purpose of financing small business start-ups, potentially supporting employment growth in localities that experience house price growth. New small businesses are considered too risky to attract either bank financing or outside equity financing. Entrepreneurs, then, must self-finance the business initially in order to establish a track record. In an influential paper, Hurst and Lusardi (2004) reported that entry rates into small business ownership are not related to household wealth, including housing wealth, for the great majority of the wealth distribution. More recent work has suggested that controlling for whether the transition to self-employment is planned reestablishes the connection between wealth and entrepreneurship. Of particular importance for our purposes is

empirical evidence that house price appreciation is positively related to small business formation (Fairlie and Krashinsky, 2012; Harding and Rosenthal, 2013). As discussed further below, this suggests that markets with inelastically supplied housing will potentially have higher rates of entrepreneurship. This is an area for future research.

15.2. HOW MORTGAGE FINANCE AFFECTS THE MARKET FOR OWNER-OCCUPIED HOUSING

Because housing is such a large purchase, representing multiple years of income for most families, households making the transition to homeownership must either make very significant consumption sacrifices in advance of the purchase or use credit. As a consequence, mortgages play a large role in the transition from renting to owning. Overall credit conditions, and their manifestation in the mortgage market, thus become an important determinant of the level of demand for owner-occupied, relative to rental, housing. In this section, we present simple models of the demand and supply sides of the housing market that are most prevalent in the current literature. Having the rudiments of a model in mind will facilitate our discussion of how various features of the mortgage contract and its treatment in tax policy affect real outcomes like prices and quantities of housing transacted and homeownership rates.

15.2.1 The demand for owner-occupied housing

We begin with the demand for owner-occupied housing units. Since everyone must consume housing services, our focus is on factors that affect the relative attractiveness of owning versus renting. Here, two features of the mortgage contract are very important. The down payment requirement determines the maximum mortgage that is available to purchase a house of given value, while the mortgage interest rate determines the annual cost of that borrowing. In addition to prevailing mortgage terms like interest rates and down payment requirements, several other costs and benefits of ownership affect the demand for owner-occupancy relative to rental units.

Public policy, particularly the tax code, treats homeownership as a special category of investment. To see how this works, it is helpful to contrast the tax treatment of owner-occupancy to that afforded to rental properties. The two forms of tenure are directly analogous as we will explain. In each case, we have a landlord who owns the property and a tenant who occupies the unit and pays rent to the landlord. This is obvious for the case of a rental property. In the case of the owner-occupied property, the landlord and the tenant roles still both exist but they now are the same person and the rental payment is implicit, but in principle, the situations are identical. Nonetheless, the treatment of these two situations under the personal income tax is quite different. The owner of the rental property will pay income taxes on the difference between the rent payments she receives and the costs of owning the property (mortgage interest, property taxes, maintenance, and

depreciation). The analogous treatment would have homeowners paying taxes on the difference between the rental income they (implicitly) pay themselves and the costs of owning the property. Instead, homeowners are allowed to exclude from income both the expenses, as a regular landlord would, and the implicit rental income they pay to themselves. In this section, we discuss the tax code's implications for the relationship between mortgage finance and housing market outcomes.

Besides the exclusion of implicit rental income, additional tax benefits to homeownership come in the forms of the exclusion of most housing-related capital gains and the deductibility of state and local property taxes from federal income taxes. These tax expenditures are expected to cost the Treasury \$80, \$57 billion, and \$34 billion in fiscal year 2015 (OMB, 2014, p. 206). While the subsidies are not directly tied to the form of finance—i.e., they do not directly affect the mortgage market—they nonetheless potentially reduce the after-tax cost of buying a home and thus will likely increase the demand for all forms of financing, including mortgage borrowing. Thus, these tax advantages interact with the preferential treatment of mortgage interest, described below, to potentially affect the ownership, form, and location of the housing stock in metropolitan areas. Poterba and Sinai (2008) offered a detailed analysis of the incidence of benefits from various elements of the tax treatment of owner-occupied housing.

The United States is not alone in providing support for owner-occupancy in its tax code. In a survey of OECD countries, Bourassa and Grigsby (2000) found that only two (the Netherlands and Sweden) tax housing capital gains. Among the sample of 24 countries surveyed by Bourassa et al. (2013), only three—Poland, the Netherlands, and Switzerland—tax the full implicit stream of rental payments. Among those 24 countries, the mortgage interest deduction is quite common, with at least partial deductibility available in 14 of the countries.

In addition to the tax benefits afforded to owner-occupancy in the United States, the rules of many other financial transactions favor owner-occupied housing, including exclusion of housing from personal bankruptcy proceedings, estate taxes, and the calculation of assets for programs such as Medicaid and student loans (Morgan et al., 2012). While these preferences are also independent of the mortgage contract, their value increases with the size and value of the house, which, as described below, is likely increased by the deductibility of mortgage interest. Because they affect the after-tax cost and benefits of owning versus renting, all of these various subsidies to owner-occupancy have the potential to affect the location of the housing demand curve.

One of the principal features of public policy that affect the net cost of credit to mortgage borrowers is the deductibility of mortgage interest from federally taxable income. Current federal tax law allows homeowners to deduct mortgage interest for loans up to \$1,000,000 used to buy, build, or improve the housing unit in which they live or one other unit. In addition, interest paid on other loans up to \$100,000 secured by these properties, for example, HELOCs, may be deducted, regardless of the purpose to which these

moneys are put. This tax expenditure is among the largest in the Internal Revenue Code and is expected to cost the Treasury \$74 billion in FY 2015 (OMB, 2014, p. 206). As a practical matter, this tax benefit reduces the cost of mortgage and HELOC interest to the borrower, with the value of the benefit depending on the borrower's marginal tax rate, so long as the borrower's deductions exceed the standard deduction threshold. For a borrower in the top bracket who faces a 35% marginal rate, an additional \$100 of mortgage interest reduces tax liability by \$35, while a borrower in the 15% bracket would receive only \$15. In addition to this federal income tax preference, 34 states also offer at least partial deductibility of mortgage interest. Hilber and Turner (2014) document the extent of these state preferences, which vary substantially over states.

The favorable treatment of mortgage interest in the tax code is a benefit of debt-financed owner-occupancy that has attracted much attention from housing and public finance economists. In addition to providing incentives for home buyers with income tax liabilities to borrow for their home purchases, thereby increasing the demand for mortgage credit, these reductions in the cost of borrowing for residential real estate have the effect of shifting the housing demand curve upward. Reduced after-tax interest costs reduce the user cost of housing, making debt-financed owner-occupied housing more attractive relative to other goods. Moreover, larger houses (mortgages) are relatively more attractive since the dollar value of the deduction rises with the interest paid. As we will discuss, the economic incidence of these tax benefits depends importantly on the supply elasticity in the local housing market.

Other features of the mortgage contract itself may also act as housing demand curve shifters. For example, reducing the down payment requirement from 20% of the purchase price to 10% means that buyers can either purchase the same home with a smaller amount of savings or buy a larger (more expensive) home with the same down payment.

This effect is amplified at low mortgage interest rates. Himmelberg et al. (2005) demonstrated the importance of mortgage interest rates in affecting the location of the housing demand curve, emphasizing the nonlinearity of the relationship between interest rates and prices. Glaeser et al. (2013) examined virtually all elements of the mortgage, from interest rates to down payment requirements, to explore the effect of these features on housing demand.

An additional benefit to homeownership comes through the credit market. Liquidity-constrained households may prefer to borrow to smooth consumption. For example, young people may wish to consume part of their (higher) permanent income today rather than to wait until they actually get the pay raise they expect. Access to, and the cost of, credit is a crucial determinant of the ability of such households to smooth consumption. Unsecured credit—collateralized only by the borrower's promise to repay the debt—offers the potential to solve this problem, but given the existence of a default (bankruptcy) option for the borrower, the lender is uncertain that the loan amount will be repaid. In equilibrium, therefore, unsecured lending reflects the cost to the lender of the borrower's

default option (Chatterjee et al., 2007). Other than housing equity, household access to secured credit is quite limited; it is primarily available to finance the purchase of certain specific consumer durables, like automobiles and, in some cases, large appliances.

Because HELOCs and CESs are secured by liens on real property, lenders are generally willing to provide this credit at interest rates considerably below that available for unsecured products like credit cards (Chatterjee et al., 2007). In addition to the interest rate discount that junior liens carry, interest paid on junior liens, like that on first liens, is exempt from federal income taxes up to a high limit. Owners will have access to this credit, while renters will not; owners of larger homes will be able to borrow more than those with smaller homes. In this way, as well as in the fact that the value of the mortgage interest tax deduction increases with income, these aspects of housing finance will be related to inequality in urban areas.

15.2.2 The effect of demand shifts on price and quantity

We now turn our attention to the supply side of the housing market, which will allow us to complete our discussion of how demand shifts driven by changes in mortgage finance and its after-tax cost will affect prices and quantities in local areas.

Housing's durability has important implications for the way that demand changes, and thus, features of housing finance will affect prices and quantities. In particular, housing supply is nearly completely inelastic at the current stock of housing for prices below replacement costs (Glaeser and Gyourko, 2005). That is, once units are built, downward shifts in demand have little effect on the number of units in place, since those units depreciate only slowly.¹¹ Instead, downward shifts in the demand for housing will be reflected largely in (reduced) prices for housing.

Reductions in demand have little effect on market-clearing quantities in local housing markets, but markets can differ in how upward shifts in demand affect quantities. Glaeser et al. (2008) explored how shifts in housing demand manifest themselves in different markets, depending on the supply elasticity in each market. Their model predictions are that any increase in demand, whatever the cause, will result in higher prices but only modest or negligible quantity effects in markets where housing is inelastically supplied. In contrast, in markets with relatively elastic supply, demand increases should result in new residential investment and less of a price response.

Figure 15.4 depicts the Glaeser et al. view of the housing market during a boom phase, or one in which the cost or availability of mortgage credit becomes more favorable. Two possible kinds of markets are simultaneously depicted in the figure. Initially, demand is represented by demand curve D_0 , and the built housing stock is H_0 . Equilibrium is shown

¹¹ The rate of physical depreciation of housing is an area that is underresearched. A complication in estimating the depreciation rate is accounting for maintenance efforts by the owners. We discuss maintenance later in the chapter. See Harding et al. (2007) for an extensive discussion and estimation using AHS data.

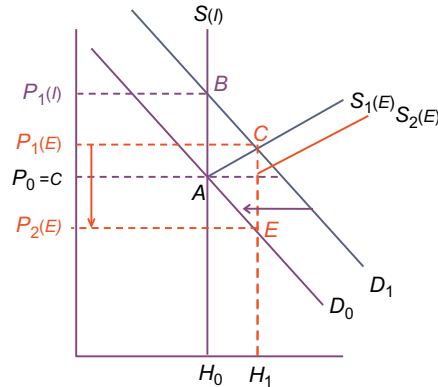


Figure 15.4 Housing market demand and supply.

as point A , with prices $P_0 = c$, where c is the marginal cost of construction. Now consider an expansion of credit or a reduction in the cost of mortgage finance that shifts the demand curve up to D_1 . In a market with perfectly inelastic supply, represented by $S(I)$, the new equilibrium is given by point B : there is no supply response and prices rise sharply, to $P_1(I)$.¹² In contrast, in a market where supply elasticity is positive, new building is feasible, and the supply schedule is assumed to be upward sloping as depicted by the segment labeled $S_1(E)$. In the new equilibrium in this elastic market, point C , prices rise to $P_1(E)$ and quantity supplied increases to H_1 .

It is worthwhile to reiterate that the durability of housing produces stark asymmetries in the effects of demand shifts on local housing markets. Continuing our previous example, if mortgage conditions change such that demand shifts back to its original position, there will be different effects in these two markets. The inelastic market, still at supply H_0 , reverts back to point A , the original equilibrium. In the elastically supplied market, however, the credit boom has induced an increase in built supply to H_1 , and the downward shift in demand reduces prices to $P_2(E)$ —a level lower than the preboom equilibrium and below replacement cost. As fundamental demand begins to expand in the elastic supply market, prices will adjust upward, but there will initially be no new building activity. Once prices have recovered to the replacement cost, new supply will again be added to the market. Overbuilding to the extent that it occurs has important consequences for local housing markets.

An extension of the Glaeser et al. view of durable housing supply is provided in [Figure 15.5](#). Following [Goodman \(2013\)](#), we allow the possibility for owners and landlords to decide to abandon property. Abandoned property can lead to vandalism and stripping of any valuable materials from the property. This rapid deterioration in the

¹² [Gyourko and Molloy \(2014\)](#) explained how local government regulations can reduce the elasticity of housing supply.

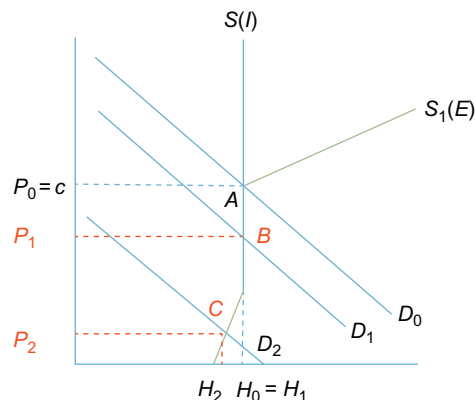


Figure 15.5 Housing market demand and supply—with abandonment.

quality of the property can result in local governments deciding to demolish the structure. In this way, some elasticity is reintroduced to housing supply at very low prices.

To explore this possibility, consider the case where housing demand declines from D_0 to D_1 . The price of housing declines from P_1 to P_2 . Assume, however, that at P_2 , the rental value of the property is sufficient to cover the variable cost to keeping the property. In this case, housing supply remains unaffected and all of the adjustment to the adverse demand shock occurs through the price of housing. Now assume that the adverse demand shock is more severe and demand declines from D_0 to D_2 . At the existing supply of housing, H_0 , the price of housing would decline by enough to push the rental value below the variable cost for many landlords. These landlords face a “shutdown” decision much like managers of a firm. If the rental value does not cover the variable cost of having a tenant in the property, then the landlord will choose instead to leave the property vacant. If demand is not expected to improve, then this will lead to the property being abandoned. This shutdown decision reintroduces some elasticity of supply at very low prices that reflects the distribution of these variable costs of utilizing a property.¹³ This acts to reduce the downward pressure on prices from the adverse demand shock. Goodman (2013) documented that this segment of the housing supply curve has been important for cities such as Detroit, which has experienced a 34% decline in its housing supply from 1970 to 2010.

The durability of housing thus implies that credit cycles will have real, long-lived implications that will differ across metropolitan areas depending on their supply elasticity. In particular, markets where housing is elastically supplied (on the upside) will be affected by a credit boom and bust through both a cycle of residential investment activity and

¹³ The degree of this elasticity depends among other factors on the length of time captured by the supply curve. The longer the time period, the greater the elasticity all else the same.

consumption swings related to housing wealth effects. In contrast, inelastic markets will see credit conditions reflected almost exclusively in prices and therefore consumption-related wealth effects, with fewer real effects from residential investment. Both will be adversely impacted by debt overhang problems following the bust (Mian and Sufi, 2009).

Glaeser et al. tested the predictions of this model using the proxy for housing supply elasticity developed in Saiz (2008).¹⁴ However, they noted that several of the markets that experienced the largest booms in the recent cycle have high measured supply elasticities. These markets also demonstrated little variability of prices relative to replacement costs prior to the recent cycle. While having an elastic housing supply limits the likelihood of a serious housing bubble in a local market, it clearly does not prevent one from happening.

We are now in a position to explore how the determinants of the demand for housing described above will interact with supply to generate the prices and quantities of housing in different markets. Each of the factors discussed above—the favorable tax treatment accorded to owner-occupied homes, the fact that home equity provides a ready source of collateral to finance consumption and investments, and the sheltering of home equity from various asset tests—serves to increase the attractiveness of owner-occupied housing relative to other assets and thus shift the demand for housing upward. While intuition might suggest that this will necessarily lead to higher homeownership and larger houses, it is now clear that the effects of this demand shift will depend on the *supply* side of the market as well.

As the above discussion of the supply side of the housing market suggested, the effect of demand shifts induced by changes in mortgage finance—for example, lower interest rates or down payment requirements—will depend on the elasticity of housing supply in various markets. In markets with inelastic supply, even large demand shifts will not increase the homeownership rate, but will instead be reflected entirely in higher prices. These higher prices, in turn, increase the value of the tax deductibility of the mortgage interest and capital gains exclusions, meaning that the preponderance of the benefit from these policies is concentrated in a few areas with high and rising prices and wealthy owners with high marginal tax rates (Gyourko and Sinai, 2004).

Hilber and Turner (2014) provide a detailed study of how the mortgage interest deduction, which has been a common target of tax reformers, affects homeownership rates. In all places, the availability of an income tax deduction for mortgage interest shifts the demand for housing upward, with the size of the shift depending on the owner's marginal tax rate. For owners with higher incomes and marginal tax rates, the value of the tax deduction is large, while for those with no income tax liability, there is no benefit. In places with elastic supply, these demand shifts induce more building and higher ownership rates.

¹⁴ This proxy is the percent of land within a 50 km radius area that has a slope of less than 15°.

But where supply is inelastic, upward shifts in the demand curve are capitalized into higher prices, making the down payment constraint more binding for young borrowers with lower incomes. This increase in cost reduces transitions into ownership for this important group, and Hilber and Turner find that the mortgage interest deduction actually reduces homeownership in inelastically supplied markets. They conclude that because of these offsetting effects, the mortgage interest deduction has little effect on overall homeownership.

The experience of other countries is consistent with this conclusion. Canada, the United Kingdom, and Australia, arguably the countries most comparable to the United States, all do not currently have a mortgage interest tax deduction, but their homeownership rates are very similar to the United States, in the vicinity of two-thirds.¹⁵ Moreover, the United Kingdom began phasing out its mortgage interest deduction in 1975 when the value of the deduction as a fraction of interest payments was 38% (Gibb and Whitehead, 2007). By 2000, the deduction had been completely eliminated, but the homeownership rate had increased from 50% in the 1971 Census to 69% in the 2001 Census,¹⁶ though naturally, a variety of other demographic, economic, and policy changes contributed to these outcomes. Andrews and Caldera Sánchez (2011) examined the factors driving the evolution of homeownership rates in 15 OECD countries. While their data do not allow direct estimation of the impact of tax policy on homeownership, they found that more generous mortgage interest tax subsidies are associated with a smaller impact of relaxing down payment requirements on homeownership rates. This is consistent with greater house price capitalization of mortgage-related demand shocks in countries where owned housing receives more favorable tax treatment, thus reducing housing affordability for the marginal buyer and reducing the demand shock's expansionary effect on homeownership.

As implied by the discussion above, both US research and international research have focused on the homeownership effect of public policies, rather than other interesting margins such as the location of new residential units. This is perhaps unsurprising since ownership is an explicit target of many of these policies and is relatively well measured. Nonetheless, there is a need for further research intended to improve our understanding of how these various subsidies affect the spatial organization of activities across and within cities and how they affect the kinds of housing occupied by both owners and renters. Such research would confront many obstacles, including the challenges of accurately measuring housing quality and of identifying the effect of policies that are chosen endogenously by the political system.

One study on the importance of the supply elasticity in determining how credit-induced demand shocks will be transmitted into local market outcomes is Davidoff

¹⁵ See Bourassa et al. (2013) exhibit 2 for a comparison of mortgage interest tax policies and homeownership rates in 24 countries.

¹⁶ UK Office of National Statistics; figures are for England and Wales.

(2013b). Davidoff found that supply elasticity as conventionally measured does a poor job of explaining the magnitude of the 2000s credit cycle, leaving open the question of why supposedly elastically supplied markets experienced such large booms and busts. One potential answer is found in Genesove and Han's (2013) analysis of the *intrametropolitan* effects of the boom/bust cycle of the 2000s. Genesove and Han showed that price increases during the boom phase tended to be concentrated in neighborhoods with good access to jobs, which, since such locations are limited, can be thought of as a supply constraint. During the bust, the spatial price gradient flattened, which is consistent with the largest declines in price coming from the most inelastically supplied areas. Much of the work on supply elasticities has been conducted at the metropolitan level, but this approach of studying price and quantity dynamics within cities may hold promise for improving our understanding of the effects of future changes in mortgage finance.

A major lesson of the recent boom and bust in housing markets is that even markets with fairly elastic supply experienced a very significant housing price cycle during the 2000s. This underscores the fact that we have much to learn about how these markets work, especially how mortgage underwriting, the sources of mortgage funding, and future price expectations—which will also have an important effect on demand—interact.

Securitization, which is the source of much of the funding for mortgages in the United States, may itself have spatial consequences. Haughwout et al. (2012) demonstrated that during the boom, many mortgages self-reported as owner-occupied loans in fact turned out to be taken out by borrowers with multiple first-lien mortgages, suggesting that they were instead operating as investors. This increase in demand by investors was facilitated by securitization, since downstream investors, unlike originators, do not have access to the borrower's credit report, which is useful for verifying the borrower's intentions. These loans were especially concentrated in boom markets in Arizona, California, Florida, and Nevada and were associated with larger differences between transactions and list prices. When market prices turned around and began to fall, these loans defaulted at very high rates, ultimately leading to spatially concentrated foreclosures.

15.3. THE DISTRIBUTION OF MORTGAGE CREDIT

One important way that mortgage finance can affect the urban landscape is through its effects on the extent and distribution of homeownership. While the homeownership rate had hovered for decades around 64–65%, Figure 15.1 shows a substantial rise and fall over the most recent housing cycle, notably corresponding to a period of expanding credit followed by credit tightening after 2007. Moreover, underneath the aggregate homeownership rates, there is considerable variation across geography and individual household characteristics. At the end of 2013, the aggregate homeownership rate was 65.2%, but only 50.2% for households with below-median income, 45.5% for Hispanics, 43.2% for blacks, and 36.8% when the householder is younger than age 35. Some of these

differences will reflect preferences, differences in the user cost of owner-occupied housing, and the assets, income, and credit history required to qualify for loans. However, discrimination in mortgage markets may also affect the ability to secure loans and the price paid for loans. In this section, we first focus on these issues of discrimination and then examine the distribution of mortgage credit over the life cycle.

15.3.1 Access to mortgage credit

Figure 15.6 shows large and persistent differences in homeownership rates by race and ethnicity: The black–white and Hispanic–white gap is never smaller than 25 percentage points. While other factors clearly play a role in explaining these differences, their magnitude naturally prompts the question of whether lenders are unfairly denying mortgage credit to black and Hispanic applicants. Post-Civil Rights era antidiscrimination laws¹⁷ require lenders to ignore the applicant’s race and ethnicity in making mortgage qualification decisions, regardless of whether these characteristics are good proxies for unobservable risk factors that affect the loan’s expected return. The laws also prohibit lenders from discriminating on the basis of the racial or ethnic composition of the neighborhood, a practice known as “redlining” because of the red lines that lenders allegedly drew around neighborhoods where they would not make loans.

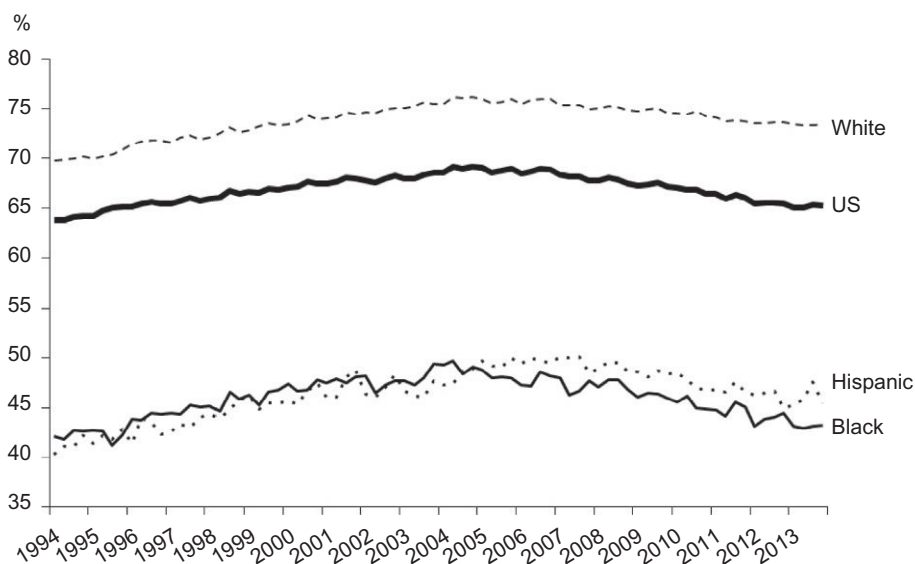


Figure 15.6 Homeownership rate by race and ethnicity. Source: US Census Bureau, Current Population Survey/Housing Vacancy Survey, Series H-111. Race and ethnicity of householder: “White” is non-Hispanic white, “Black” is non-Hispanic black, and “Hispanic” can be of any race.

¹⁷ These include the Fair Housing Act of 1968 and the Equal Credit Opportunity Act of 1974.

Earlier research on the role of individual race and ethnicity on mortgage denials was hampered by a lack of adequate data on key underwriting variables used to predict default and the cost of default. In particular, Home Mortgage Disclosure Act (HMDA) data, which included many of the necessary variables, lacked information on the applicant's credit history and the property value (thus, LTV could not be calculated). These limitations were addressed in the seminal Boston Fed study (Munnell et al., 1996) that examined HMDA data from Boston, merged with a comprehensive set of additional variables collected directly from lenders. They estimated a standard loan denial regression, including an indicator for black or Hispanic applicant. Even after extensive controls, the researchers found that loan denials were 8 percentage points higher for blacks and Hispanics. The Boston Fed made the underlying data available and their results have been subject to a great deal of scrutiny. Ross and Yinger (2002) provided a comprehensive assessment of this literature and concluded that the findings are generally robust, providing convincing evidence that Boston lenders discriminated against blacks and Hispanics in 1990.¹⁸

The research on geographic redlining has been fraught with the same inference problems as studies examining discrimination based on individual race and ethnicity: redlining cannot be isolated without information on individual credit history.¹⁹ Of the few studies that have adequate controls, there is little or no evidence that lenders systematically denied applications based on neighborhood race and ethnicity. In an extension to the Boston Fed study, Tootell (1996) estimated loan denial regressions similar to the original study but replaced the census tract dummy variables with a set of census tract characteristics. He found that a census tract's racial and ethnic composition had no statistically significant effect once individual race and ethnicity are controlled for and neither did high rates of vacancy and boarded-up property in the tract nor a large share of low-income residents. Thus, there was clearly individual-based discrimination against nonwhites in Boston, but no evidence of redlining. Hunter and Walker (1996) came to a similar conclusion, also using the Boston Fed study data.

Going beyond the prohibition on redlining, the Community Reinvestment Act of 1977 (CRA) imposes an affirmative obligation on banks (federally insured depository institutions) to help meet the credit needs of their entire service area, including low- and moderate-income neighborhoods, which often have disproportionately nonwhite populations. Bernanke (2007) explained that lending in lower-income neighborhoods can be challenging as fewer home sales and more diverse housing structures make accurate appraisals more difficult and the short or irregular credit histories of lower-income borrowers makes credit evaluations more costly. The CRA attempts to rectify market failures by inducing banks to invest in the necessary expertise to lend in underserved

¹⁸ See also Ladd (1998) for a summary.

¹⁹ See Schill and Wachter (1993) for a review of earlier studies.

neighborhoods and reduce the potential first-mover problem where no lender has an incentive to be the first to enter a new market.

Research on the effectiveness of the CRA has tended to find positive effects. For example, [Avery et al. \(2003\)](#) found that the census tracts just below the CRA income thresholds had higher homeownership rates, higher growth in owner-occupied units, and lower vacancy rates than those just above; [Apgar and Duda \(2003\)](#) also concluded that the CRA expanded mortgage access for lower-income borrowers and suggested that its coverage be expanded. However, there are critics that argue that the CRA has been ineffective, not as effective as alternative policies, or that its costs outweigh its benefits. [Barr \(2005\)](#) provided a good overview and critique of these criticisms, ultimately concluding that the CRA is a reasonably effective policy to overcome market failures and discrimination.

15.3.2 The distribution of high-cost lending

Mortgage lending practices rapidly changed in the early 2000s with a large expansion of credit, especially in the subprime market, as shown in [Figure 15.7](#). New mortgage products requiring lower down payments, lower credit scores, and possibly little or no documentation of assets and income became much more widely available, fueled by the increased securitization of subprime mortgages. But of course, these products typically came at additional cost to the borrower. Moreover, these additional costs were potentially unclear to borrowers because the loans often had relatively complex features such as initial “teaser” rates that would be followed by a large payment jump even if market

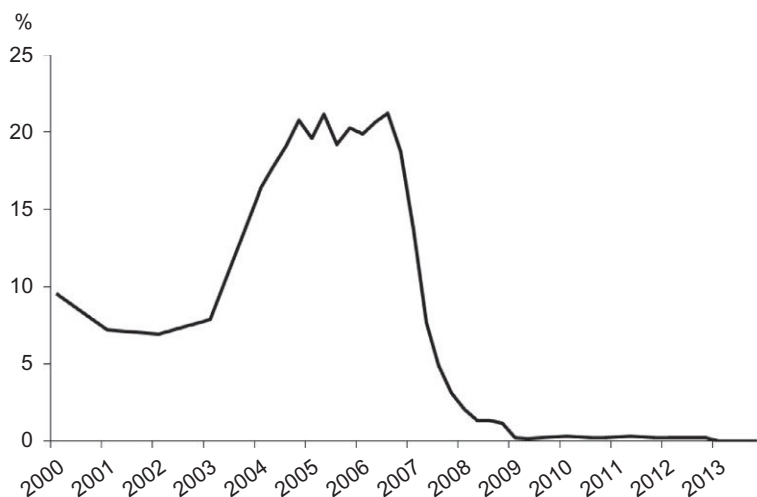


Figure 15.7 Fraction of mortgages that is subprime. Source: *Inside Mortgage Finance*, 2013. Weighted by mortgage balance.

interest rates remained constant.²⁰ This led to a heightened concern that low-income and nonwhite individuals and neighborhoods were disproportionately receiving higher cost loans, particularly those with large expected rate resets that borrowers could not possibly afford. Lenders and mortgage brokers were accused of steering borrowers to inappropriately structured mortgages with high costs. On the borrower end, the concern was not only about adequate disclosure but also that individuals with limited financial literacy may not understand how the teasers work and underestimate how much their rate could increase, or they may overestimate their ability to pay the higher reset rate or simply not consider it important due to myopia or hyperbolic discounting.

Numerous studies have documented patterns of high-cost or subprime lending by income, race, and ethnicity. For example, [Mayer and Pence \(2009\)](#) examined several data sources and found that in 2005, subprime mortgages were increasingly concentrated in zip codes with higher proportions of black and Hispanic residents, even after controlling for zip code income and credit scores. As measured by the fraction of black residents, the 90th percentile zip code had 42% more subprime loans than the corresponding median zip code, and as measured by the fraction of Hispanics, the 90th percentile zip code had 33% more subprime originations. They also suggested that subprime loans are more prevalent in locations where credit might be more difficult to obtain, such as zip codes with midlevel credit scores or counties with higher unemployment rates.

[Figure 15.8](#) provides an illustration of the geographic concentration of subprime lending in predominantly black and Hispanic neighborhoods in New York City. The shading shows the share of black or Hispanic residents in each census tract, while each dot represents the census tract location of one higher cost loan originated in 2006. As defined by HMDA, a higher cost loan is one where the annual percentage rate (APR) is greater than 3 percentage points above a Treasury of comparable yield—this is a commonly used measure of subprime mortgage.

While not considered part of the subprime market, [Pinto \(2012\)](#) argued that many FHA FRMs originated during this period displayed characteristics that were typical of subprime loans including low credit scores, low down payments, and high debt-to-income ratios, particularly those originated in zip codes where median family income was below the median family income for the metropolitan statistical area (MSA). Like their subprime counterparts, these loans experienced very high default rates, resulting in disproportionately high foreclosure rates in disadvantaged neighborhoods.

To the extent that blacks and Hispanic households with low income are more likely to be credit-constrained and less able to qualify for conventional loans, the expansion of subprime high-cost lending could be welfare improving as it enhances the ability to own a home or to withdraw home equity. However, if subprime loans were substitutes for

²⁰ [Mayer et al. \(2009\)](#) provided a good description of the structure and risk characteristics of subprime and near-prime mortgages.

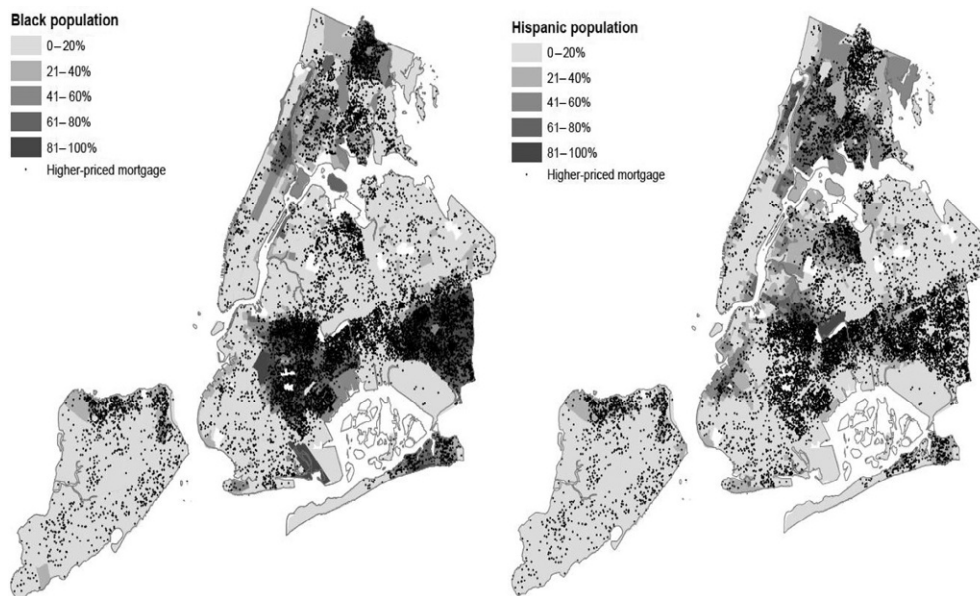


Figure 15.8 Distribution of higher cost mortgages in NYC, by census tract race and ethnicity. *Source: US Census 2000 and HMDA 2006. Each dot represents one higher-priced mortgage origination (first lien, owner-occupied home purchase), randomly located within its census tract. Shading shows the fraction of non-Hispanic blacks, and fraction Hispanics of any race among each tract's population.*

cheaper conventional loans, then the outcome is clearly of concern for equal access to credit reasons, even without considering the subsequent foreclosure crisis and its associated negative spillovers. Disentangling these two effects is a challenge because it is impossible to construct precisely the set of mortgages that a borrower could have qualified for. Furthermore, mortgages are differentiated along many dimensions. Loans with different terms, amortization and rate schedules, and prepayment penalties, for example, will have different values for borrowers depending on their expected housing duration and expectations about housing appreciation and their own future income trajectory. Without information on these individual factors, it can be difficult to conclude that one loan is necessarily expected to be more costly than another for a given borrower, even if the whole choice set is observed.

Many studies examine mortgage pricing using HMDA's APR measure, which amortizes interest payments and up-front fees over the full maturity of the loan and reports the spread over a Treasury security of comparable maturity when this spread is at least 3 percentage points. This collapsing of price into a single left-censored measure presents several problems that should be borne in mind in considering the literature. First, for a given combination of up-front fees and rising rate schedule, a borrower with a shorter expected housing duration will have lower costs, despite the same measured APR. Second, borrowers rarely hold a mortgage to its maturity. Two mortgages can have the same APR

and have quite different expected costs assuming an expected duration shorter than maturity. To the extent that nonwhites are more likely to be targeted for high up-front fees, HMDA's APR measure would not be able to uncover this form of discrimination and would underestimate disparities. Third, leaving aside any up-front costs, the APR concept may also be problematic with hybrid adjustable rate mortgages (ARMs) where a substantial portion of borrowers prepay at the first reset (many loans had prepayment penalties that expired at this point). A lender can thus tweak the rate adjustment formula to achieve any APR without significantly impacting the likely payment stream generated from the mortgage. Finally, the left censoring may prevent discrimination of borrowers with better credit from being uncovered as they are less likely to have rates above the left-censoring threshold.

Analyses of the post-2000 data that rely solely on HMDA data (for example, [Avery et al., 2006](#)) tend to find substantial adverse pricing effects for blacks and Hispanics, but they are severely hampered by a limited set of explanatory variables, including no information on credit scores or LTV. [Bocian et al. \(2008\)](#) overcame this problem by merging HMDA with a large proprietary subprime loan dataset containing much more information on borrower risk attributes. They run their analysis separately for subprime 30-year FRMs and the most popular ARM, the 2/28 (a hybrid ARM with an initial 2-year teaser rate followed by rate resets at 6-month intervals for the remaining 28 years). They further segmented these two groups by whether they had prepayment penalties and by purchase versus refinance loans. In general, their analyses show that among home purchase loans with prepayment penalties, black and Hispanic borrowers were significantly more likely to receive APRs that exceeded HMDA's reporting threshold of 3 percentage points above Treasuries and that these disparities were greater for FRMs compared with 2/28 ARMs. However, there was little significant difference for either FRMs or ARMs without prepayment penalties or for refinance loans. Bocian et al. did not estimate the *amount* by which race and ethnicity increase the APR and implied that the left censoring makes it impossible to do so. Their focus on the subprime market leaves open the broader question of whether some borrowers were steered into subprime mortgages when they could have qualified for cheaper conventional loans.

[Courchane \(2007\)](#) was able to incorporate both prime and subprime loans by using a (nonrepresentative) sample from 22 lenders. These proprietary data are not constrained by HMDA's left censoring of the APR and contain the critical underwriting variables of debt-to-income ratio, credit score, and LTV. Courchane examined 2004–2005 originations to estimate the probability of taking out a subprime mortgage and the determinants of APR conditional on receiving a subprime or prime mortgage using an endogenous switching framework. The results suggest that virtually all of the race–ethnicity difference in the use of the subprime market and the vast majority of the black–white and Hispanic–white APR gap can be explained by observables that are appropriately used in underwriting and pricing. The remaining unexplained gap in

APR is economically small,²¹ and the analyses suggest that they primarily come from within the prime or subprime market, rather than from selection across the two markets.

More recently, [Bayer et al \(2014\)](#) link 2004–2008 HMDA data to credit reports and to public property and mortgage deeds in seven cities. Their merged data allow them to control for a rich set of risk attributes including the presence of a junior lien and to capture both the prime and subprime markets. However, they remain hampered by HMDA's left-censored APR as their measure of high-cost pricing and only model the likelihood that a borrower will receive a loan with an APR that exceeds HMDA's threshold. Bayer et al. find significant and substantial unexplained disparities, with black, Hispanic, and older borrowers more likely receiving high APR loans. In particular, they find the largest difference exists for black home buyers across the credit score spectrum, especially in counties where recent black home buyers are less likely to have some college education. For Hispanic borrowers and older borrowers, the difference is primarily concentrated among borrowers with low credit scores or high LTVs. Bayer et al. also conclude that substantial amounts of the estimated disparities are due to sorting across lenders as the disparities are substantially reduced when lender fixed effects are included in the models.

Another strategy that has been taken by several studies is to examine some of the components of APR, rather than HMDA's APR itself. This method necessarily has to sidestep the issue of selection and steering by focusing more narrowly on specific products to examine whether lenders price a given product differently for different borrowers, but is able to generate an estimate of the size of any pricing disparity. [Haughwout et al. \(2009\)](#) used data from the servicing records of securitized nonprime mortgages merged with HMDA and focused on 2/28 ARMs. In contrast to much of the previous literature, they found that after controls for risk characteristics and neighborhood composition, black and Hispanic borrowers get slightly more favorable terms, although the sizes of these effects are economically tiny.²² On the other hand, Asian borrowers paid slightly higher initial rates and reset margins. They also found that loans were cheaper in zip codes with a higher percentage of Asian, black, and Hispanic residents, as well as in counties with higher unemployment rates, consistent with the higher concentrations of subprime lending in these locations found by [Mayer and Pence \(2009\)](#). One drawback of their data is that they do not include information on any points or fees that the borrower may have paid at origination, so it remains possible that the estimated lower interest rates and reset margins are due to higher fees having been paid or that, in fact, higher fees were disproportionately charged to nonwhite borrowers, even after accounting for slightly lower rates and margins. And of course, these analyses leave open the broader question of selection into products.

²¹ The APR was 0.09–0.10 percentage points higher for blacks and 0.08–0.11 higher for Hispanics, compared with white borrowers.

²² Initial mortgage rates are about 0.025 percentage points lower for blacks and Hispanics than the average of 7.3%, and reset margins are 0.017–0.05 percentage points lower than the average of 5.9%.

Ghent et al. (2014) use similar data to focus on securitized nonprime originations in California and Florida in 2005. They examine the seven most popular nonprime mortgage products and a category for the remainder, but like Haughwout et al. (2009), they conduct their analyses for each product separately and do not consider steering or selection either among the various products or with the prime market, and they cannot examine up-front points and fees. Their empirical strategy differs from Haughwout et al. in that they also control for possible differences in default and prepayment probabilities in their pricing models. Ghent et al.'s findings suggest adverse pricing for black and Hispanic borrowers, and in neighborhoods with a higher share of nonwhite residents, but the magnitude of these effects is relatively small.²³ Their results also indicate that these disparities cannot be entirely explained by the (significant) effect of individual or neighborhood race and ethnicity on the 2-year probabilities of default or prepayment. It is important to emphasize that this form of statistical discrimination is illegal, but nonetheless, the analysis is useful for understanding the source of any disparities. These findings suggest that the disparities go beyond this particular illegal form of statistical discrimination.

Interestingly, Ghent et al. find virtually no differential pricing for refinance mortgages. Because refinance borrowers already have a mortgage and are thus more knowledgeable with respect to mortgage markets, the authors argue that this finding suggests that the adverse pricing is "not due to discrimination on the part of lenders *per se*." Rather, they argue it is due to nonwhite home buyers searching less intensively or effectively for the best possible rate compared with similar white home buyers, perhaps because nonwhites are less likely to benefit from community or intergenerational transfers of mortgage market knowledge. It is worth noting though that the more knowledgeable refinance borrowers are also less likely to be susceptible to aggressive marketing tactics and steering, and thus, the lack of an adverse pricing effect for refinance borrowers could simply be because they are more able to resist the possibly discriminatory actions of mortgage brokers or lenders. Finally, if there are characteristics of the borrowers that are unobserved to the econometrician but observable to loan underwriters, then borrowers seeking to refinance a mortgage are a selected sample relative to borrowers seeking to purchase their first home. Thus, pricing differentials observed for purchase mortgages could reflect sample selection.

Some research has focused specifically on the role of mortgage brokers in determining the price paid by borrowers. Brokers were able to receive compensation directly in the form of cash from the borrower (points), as well as indirectly via a commission from the

²³ For example, for the 30-year ARM (the most product), black and Hispanic borrowers faced interest rates that were, respectively, 0.12 and 0.29 percentage points higher than other borrowers, and a 10 percentage point increase in the neighborhood share of blacks or Hispanics is associated with, at worst, a 0.014 percentage point increase in rates.

lender to the broker (the yield spread premium), which results in a higher interest rate for the borrower. Woodward and Hall (2012) examined the sum of these two sources of broker commissions using data on a sample of FHA fixed-rate home purchase loans originated in 2001. In their data, black and Latino borrowers paid higher total broker commissions than white borrowers with similar loan amounts and credit scores, while borrowers from census tracts with higher educational attainment paid lower commissions than similar borrowers elsewhere. Their results suggest that shopping from too few brokers resulted in borrowers sacrificing at least \$1000 and that borrowers who compensated brokers with both cash and a lender commission paid twice as much as similar borrowers who paid no cash. They argued that the primary reason for this seemingly irrational behavior is borrower confusion—borrowers may misunderstand the broker’s role to be helping them search for the “best” mortgage, and they may think that points and the loan’s interest rate are independent rather than a trade-off. Woodward and Hall’s conclusion is consistent with Courchane et al. (2004) who analyzed data from a survey of prime and subprime borrowers in 2001. They found that borrowers who engaged in less search activity and had less mortgage market knowledge were more likely to get a subprime loan, even after controlling for underwriting variables such as LTV and credit score.

One conclusion from this strand of literature is that borrowers would benefit substantially by more search activity, including shopping from additional brokers. But Hanson et al. (2013) also provided some evidence that mortgage brokers treat borrowers differently depending on their race and their credit score. They conducted a matched-pair field experiment in which brokers were e-mailed requests for assistance with obtaining a mortgage from fictitious borrowers who varied by credit score and by race.²⁴ Hanson et al. found a small but significant difference in the likelihood of responding to black clients compared with white clients.²⁵ Larger response differences were found by credit score, and these exacerbated differences in the differential response by race. Furthermore, even among brokers that responded to both races, the content of the replies tended to offer more details and used more friendly language (as judged by an external review panel), although overt discriminatory actions such as offering less favorable terms or steering into a product was rare.

Any differential treatment by mortgage brokers should be considered in combination with the behavior of other actors in the residential real estate market. The Department of Housing and Urban Development (HUD) has conducted a series of paired-testing studies beginning in 1977 to monitor discrimination by real estate agents. In the most recent,

²⁴ Race was signaled through the use of names with a high likelihood of association with one race. See Ross et al. (2008) for an earlier example of matched-pair testing in the mortgage application process using live testers in 2000, which found that nonwhites received less information and assistance than comparable whites in Chicago, but not in Los Angeles.

²⁵ On net, 1.9% of brokers do not respond to inquiries from black clients but respond to white clients.

Turner et al. (2013) found that black and Asian home buyers are told about and shown significantly fewer available homes than equally qualified whites.²⁶ This differential treatment by agents not only constrains choices but also raises the cost of housing search and exacerbates any differential treatment or differences in search costs experienced in the mortgage application process.

15.3.3 Mortgage finance over the life cycle

The simple model of the demand for owner-occupied housing presented earlier in this chapter focused on the user cost of housing. Standard economic theory posits that in the absence of liquidity constraints, households choose to consume goods (including housing) to maximize their utility in each period subject to their lifetime budget constraint. Preferences and permanent income will thus determine the amount of housing that a household demands, and the user cost of housing will determine whether the household rents or owns. But of course, liquidity constraints are real and important, in terms of both access to credit and the ability to liquidate a large lumpy asset. The structure of mortgage finance affects the tenure decision and the amount of housing consumed over the life cycle, with implications for the urban landscape.

There is a clear pattern of increasing homeownership rates as households age that is mirrored by declines in mortgage holding, as shown in Figure 15.9. Below, we discuss the

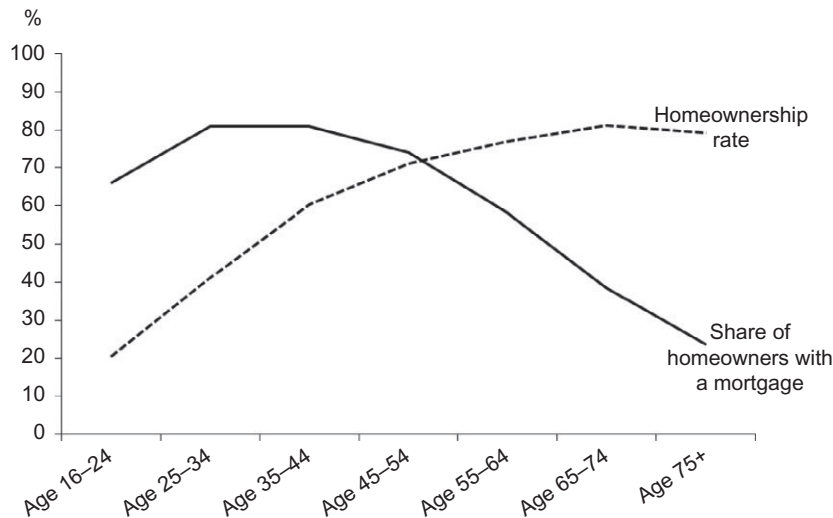


Figure 15.9 Homeownership and mortgages over the life cycle. Source: Current Population Survey, March 2013.

²⁶ Black and Asian homebuyers who contact agents about recently advertised homes for sale learn about 17.0% and 15.5% fewer available homes, respectively, and are shown 17.7% and 18.8% fewer homes, respectively. No significant difference was found for Hispanic homebuyers.

implications of mortgage finance for periods in the life cycle where current income is least likely to equal permanent income. At younger ages, borrowing constraints are most important, while at older ages, the ability to draw down housing equity is the biggest concern.

15.3.3.1 Transitions to homeownership

Even in the absence of liquidity constraints, we would expect younger households to have lower rates of homeownership for several reasons. First, households early in their career will tend to have lower marginal tax rates, which increases the user cost of owner-occupied housing, all else equal. Second, younger households, particularly those who are single and childless, have more flexibility in moving for labor market opportunities and possibly a higher propensity to search for a residential location with their preferred mix of services and amenities. This preference for greater mobility means that they are less willing to incur the substantial transactions costs associated with buying and selling a home, compared with moving between rental homes. Third, in many parts of the United States, owned homes and rental homes are fundamentally different in terms of their physical attributes, surrounding neighborhood, and access to local public goods. Rental options for detached single-family homes with private outdoor space and amenities such as high performing public schools and low crime rates are often quite limited in some jurisdictions, partly due to zoning regulations. These provide a motive for owning versus renting, beyond user cost considerations, that will be less important for younger households given their preferences.

That said, credit constraints arising from the structure of mortgage finance undoubtedly constrain some households in the timing of making a transition from rental to owner-occupied housing. To qualify for a mortgage, borrowers generally need to demonstrate a credit history that signals a willingness to repay debts and a steady income stream that signals the ability to pay; households early in their careers are less likely to possess either. Younger households are also more likely to have outstanding student loans, reducing the ability to borrow further for housing purchases. But it is arguably the down payment requirement that poses the greatest hurdle for renters who aspire to be owners.

Surveys of renters have consistently reported saving for a down payment plus closing costs to be a major obstacle to homeownership.²⁷ The Census Bureau's Housing Affordability Index uses data from the Survey of Income and Program Participation (SIPP) to capture the ability of households to purchase a modestly priced home in their area, based on their income, assets, and debts, and the assumption of a 30-year conventional mortgage with a 5% down payment (see [Wilson and Callis, 2013](#)).²⁸ The SIPP lacks credit

²⁷ [National Association of Realtors \(2013\)](#).

²⁸ A "modestly priced home" is defined as the 25th percentile of owner-occupied homes in the household's metropolitan area or within the household's state.

history information so this affordability measure cannot capture borrowing constraints arising from poor credit history. In 2009, the most recent year available, 93% of renters could not afford to buy a modestly priced home, and among them, one in four had sufficient income to qualify for a mortgage but not enough cash for a down payment.²⁹

For many young households, saving for a down payment will likely involve less consumption of both housing and other goods than would have occurred otherwise. Engelhardt (1994) used the Panel Study of Income Dynamics (PSID) to estimate the sacrifice associated with saving for a down payment and found that real food consumption growth is 10% lower during periods when a household is saving for a first home, compared to after the home purchase.

There is also good evidence that transfers from family members alleviate down payment constraints and accelerate the transition to homeownership. For example, Cox and Jappelli (1990) found that intergenerational transfers are intentionally directed towards younger generations facing liquidity constraints. Engelhardt and Mayer (1998) found that transfers from parents lead to earlier purchases of more expensive homes and that the prevalence of family help is substantial: 22% of first-time home buyers receive a down payment gift and the average gift is more than half the down payment.

These intergenerational transfers can serve to perpetuate wealth concentration and homeownership disparities across income, race, and ethnicity.³⁰ Homeownership is widely recognized as a tool for wealth building among lower-income households because the monthly payment on an amortizing mortgage serves as a form of forced saving. Moreover, the effect of leverage means that a mortgage borrower gets the benefit of the entire increase in the home's value despite only putting down a fractional equity stake (the down payment). Herbert et al. (2014) argue that these wealth accumulation benefits of homeownership for lower-income households remain intact despite the recent housing boom and bust. Down payment constraints may thus exacerbate wealth concentration by postponing wealth building for households that do not have access to family gifts. The problem is even worse when we consider the possibility that the accumulated home equity is further used in wealth creation, as capital for small businesses or to finance human capital accumulation of younger generations.³¹

Another way of viewing the borrowing constraint is in terms of the all-or-nothing nature of the homeownership decision. Housing costs are a large share of most families' budgets, and at present, there is limited scope for dividing housing into part-rent and part-own. Typically, mortgages cover the entire house and thus down payment

²⁹ An additional 74% had both insufficient income and insufficient cash, and 2% had insufficient income only.

³⁰ Hilber and Liu (2008) provided evidence that parental transfers explain part of the black-white gap in homeownership.

³¹ Lovenheim (2011) found that housing wealth has a substantial positive effect on college enrollment, particularly for low- and middle-income families.

requirements are substantial relative to incomes. However, housing tenure options that fall in between rental housing and traditional homeownership do exist and are often collectively known as shared equity homeownership. In this alternate form of ownership, the resident consumes the entirety of housing services, but is a partial owner of the underlying housing asset and takes a fraction of any appreciation or depreciation.

Existing shared equity programs generally rely on an institution to provide the remaining share of equity. [Lubell \(2014\)](#) gives an overview and assessment of models that focus on lower-income households, typically with a local government or a not-for-profit institution providing the shared equity. Some universities operate shared equity housing programs for their employees in the form of shared appreciation mortgages.³² The structure of these mortgages varies, but they generally offer the borrower reduced or deferred interest payments in exchange for a share of the appreciation.

In contrast to these existing models, [Caplin et al. \(1997a\)](#) had devised a system of shared ownership that relies on an investor market for the nonresident equity share. In these “housing partnerships,” the homeowner and resident (the managing partner) and a private investor (the limited partner) share the purchase of the home and the subsequent proceeds from its sale. The amount of borrowing required by the homeowner is scaled back proportionate to their share of the partnership and thus the mortgage (both down payment and monthly payments) becomes much more affordable. Caplin et al. envisioned a secondary market for limited partnerships, analogous to those for regular mortgages.

15.3.3.2 Mortgage finance at older ages

It is well known that in coming decades, the US population is likely to age considerably due to increases in life expectancy and the aging of the baby boom generation.³³ This aging makes it particularly important to consider the impact of mortgage finance on older adults, as it may shape their choice of location and consequently the kinds of community support and services that should be provided. There is a widespread belief among policy makers and advocates that aging in place—living in one’s home for as long as possible—is a desirable outcome that benefits older adults and their communities and that yields individual and public sector cost savings.³⁴ Appropriate mortgage finance instruments and institutions may have a role to play in increasing older households’ ability to age in place.

At older ages, an important issue for homeowners is how they can draw down their accumulated home equity or convert it into a stream of income for consumption during retirement. As previously shown in [Figure 15.9](#), homeownership rates increase with age,

³² For example, Harvard, NYU, and Stanford offer these programs to some employees.

³³ According to projections, the population aged 65 and above is expected to more than double between 2012 and 2060. US Census Bureau: 2012 National Population Projections.

³⁴ See [HUD \(2013b\)](#) for a summary of evidence.

reaching over 80% for householders aged 65 and above, while the share of homeowners with a mortgage is monotonically declining with age. Moreover, home equity is the primary component of nonpension wealth for most seniors.³⁵ But homes are lumpy and illiquid assets, which complicates matters. Homeowners who want to reduce their housing equity have several options. Selling the home and moving to another of lesser value or to rental housing is an obvious solution to extracting home equity, but it requires the homeowner to move. This likely involves substantial psychic costs: a widely reported 2010 AARP survey found that three quarters of respondents aged 50 or above strongly agreed with the statement “what I’d really like to do is stay in my current residence for as long as possible.”³⁶

Allowing the home to depreciate by reducing home maintenance is another method of consuming home equity. But unlike the temporary deferral of home maintenance for consumption smoothing purposes, a persistent reduction is unlikely to be optimal. Davidoff (2006), using data from the AHS, found that homeowners aged 75 or more spend significantly less on routine maintenance and have a much lower propensity to undertake major repairs, resulting in reduced housing appreciation that greatly exceeds the savings from undermaintenance. This difference could be interpreted as the cost associated with extracting home equity using this approach.

Taking on greater housing debt using additional mortgages, HELOCs, or via a cash-out refinance would be the straightforward way to use mortgage finance to reduce home equity without having to move. But for retired homeowners, additional monthly payments may pose a qualification constraint for borrowing,³⁷ and uncertainty regarding medical costs at older ages may lessen the attractiveness of higher monthly obligations during retirement. HELOCs have the additional problem of balloon payments that the borrower may not be able to refinance when the line of credit ends.

Reverse mortgages have the advantage of allowing homeowners to withdraw equity without having to make monthly mortgage payments, and the loan has to be repaid only when the borrower moves, sells the home, or dies. The FHA-insured Home Equity Conversion Mortgage (HECM) currently accounts for nearly all reverse mortgages in the United States. HECMs allow borrowers aged 62 and above to access their home equity as a line of credit, as a term or lifetime annuity, as a lump sum, or in some combination of these options. These loans are nonrecourse to the borrower and any heirs. The sizable insurance premiums are intended to cover the risk of negative equity at termination, either because the borrower lived longer than actuarial expectations or because home prices declined.

³⁵ Poterba et al. (2011).

³⁶ Keenan (2010).

³⁷ In particular, the declining popularity of defined benefit pensions coupled with low rates of annuitization may have led to less pension income, as opposed to wealth, in retirement.

Given their advantages and the potentially large numbers of people that could qualify and benefit from them, a long-standing puzzle had been why reverse mortgages remained relatively unpopular. There is a substantial literature on whether the elderly want to, or should, consume their home equity.³⁸ Besides pecuniary attachment to a home that may lead to a preference for leaving housing, as opposed to other assets, as a bequest, and the favorable treatment of home equity by some social programs,³⁹ the literature has emphasized the usefulness of a nonannuitized stock of wealth, such as housing, as a source of precautionary saving.⁴⁰ But while older households may be reluctant to draw down home equity because of the insurance it provides, when adverse events do in fact occur, it is important for them to have a vehicle by which they can access that equity.

Since the mid-2000s, demand for reverse mortgages has increased dramatically, though it has subsequently fallen and, while originations are still much higher than in 2000, they remain low. [Davidoff \(2013a\)](#) argued that, under a set of reasonable assumptions, HECMs historically offered borrowers favorable pricing, and thus, weak demand cannot be attributed to their high cost. Several studies indicate that the growth in demand in the 2000s mirrored the credit expansion for other forms of housing debt in that it was at least partly driven by house price growth. [Shan \(2011\)](#) provided a descriptive analysis of trends from 2003 to 2007 and concluded that house price increases account for about one-third of the growth in the reverse mortgage market. More recently, [Haurin et al. \(2014\)](#) examine state-level variations in the take-up rate of HECMs from 2000 to 2011 and find that states where real house prices were more volatile relative to historic trends had substantially higher take-up rates. They argue that this behavior is consistent with homeowners anticipating future reductions in house prices and locking in their equity gains.

The [Consumer Financial Protection Bureau \(2012\)](#) reported that the age distribution of HECM borrowers has become younger and that borrowers have more debt and are increasingly taking the full amount for which they qualify as a lump sum up-front payment (73% of HECM borrowers in 2011). They suggested that borrowers are increasingly using HECMs to refinance traditional mortgages without having to make monthly payments; essentially, they are using their existing home equity to service the debt. This puts them at greater risk for mobility lock-in due to negative equity, compared to a traditional mortgage product, and of course, they now have less precautionary saving for unexpected expenses or to finance a future move. While there is no labor market impact from lock-in for retired homeowners, the death of a spouse or an adverse health

³⁸ See [Poterba et al. \(2011\)](#) for a summary.

³⁹ Provisions vary by state, but the primary residence is often excluded in asset tests for Medicaid and Supplemental Security Income eligibility.

⁴⁰ For example, [Davidoff \(2010\)](#) argued that households tap into their housing equity when they need long-term care and that this explains the low demand for long-term care insurance.

event that limits daily activities may make moving to another residence a preferred strategy. But for individual borrowers, these risks need to be balanced against the benefit of additional cash flow and the ability to remain in their home indefinitely. While there is some limited survey evidence on how and why borrowers use reverse mortgages, this is an area where information and research are currently lacking.

There is no research that we are aware of on maintenance and home investment by reverse mortgage borrowers. Under the HECM terms, borrowers are required to keep the home in good repair.⁴¹ The findings of [Eriksen et al. \(2013\)](#) suggest that older homeowners should be investing more in home maintenance and improvements as they age in order to prevent accidents that precipitate large medical costs.⁴² Since reverse mortgages can be used to help fund home improvements, they have the potential to help adapt the existing housing stock to an increasingly aged population.

The HECM insurance fund has sustained heavy losses since the housing market passed its peak, because of negative equity at loan termination.⁴³ Moreover, by February 2012, nearly 10% of HECMs were in default for failure to pay property taxes or homeowner's insurance, placing many homeowners at risk of foreclosure. The two largest originators (with 36% market share between them) exited in 2011, with one of them citing reputational concerns from foreclosing on seniors ([Consumer Financial Protection Bureau, 2012](#)). As a response to these problems, HUD has recently limited the ability to take large up-front payments and has introduced new underwriting criteria based on income and creditworthiness to assess HECM borrowers' ability to meet tax and insurance obligations. To date, research on the determinants of reverse mortgage default has been hampered by the lack of appropriate data as risk characteristics such as credit score, debt, and income were not used in the loan approval process.⁴⁴

Problems with the HECM program may reflect the historic up-and-down swing that the housing market has experienced over the last decade and not its long-run sustainability. However, that sustainability, or lack of sustainability, may be an important factor in shaping communities in the future. With HECM's dominant position in the reverse mortgage market, its continued functioning may be an important factor affecting homeowners' ability to draw on home equity while aging in place.

⁴¹ If the home falls into bad repair and the borrower does not make repairs when requested, the lender has the right to foreclose.

⁴² Their estimates suggest that each dollar spent on home safety and accessibility features is associated with a 93 cent reduction in medical costs from fewer nonfatal falls and, that for those 75 and older, the reduction in medical costs far exceeds a dollar-for-dollar return. These estimates only count medical and not any psychic costs or the value of any formal or informal postacute care.

⁴³ [Integrated Financial Engineering \(2012\)](#).

⁴⁴ Preliminary findings from [Moulton et al. \(2014\)](#) suggest that credit score, prior delinquencies, and large up-front payments are important predictors. [Davidoff \(2014\)](#) argues that selection on home price movements and the demographics of HECM borrowers explain roughly half of the poor performance in HECM loans.

15.4. NEGATIVE EQUITY

The significant house price declines during the housing bust created an unprecedented level of negative equity. Figure 15.10 shows the shares of mortgages that were in negative equity over time. CoreLogic reported that in the fourth quarter of 2009, a total of 11.3 million mortgages or 24% of borrowers were in negative equity. Data from the American Community Survey indicate that roughly two-thirds of homeowners have a mortgage. This implies that at the end of 2009, nearly 16% of homeowners were in negative equity. Since 2009, the combination of house price increases in most markets, debt amortization, and foreclosures has slowly reduced the extent of negative equity. CoreLogic reported for the third quarter of 2013 that the number of mortgages in negative equity was down to 6.4 million.

Given the significant magnitude of negative equity that was created as a result of the housing bust, it is important to explore possible consequences of this negative equity for local housing markets. Three important areas of inquiry are implications for strategic default, the turnover rate of the housing stock, and the degree to which homes in negative equity are being adequately maintained. Strategic default clearly affects foreclosure rates, a topic we will discuss later in this chapter. The turnover rate is important in that it determines the degree to which households are able to sort across houses over time. Turnover also helps to promote price discovery. Furthermore, there tends to be significant expenditures made by households when they move into a home.⁴⁵ In addition, one

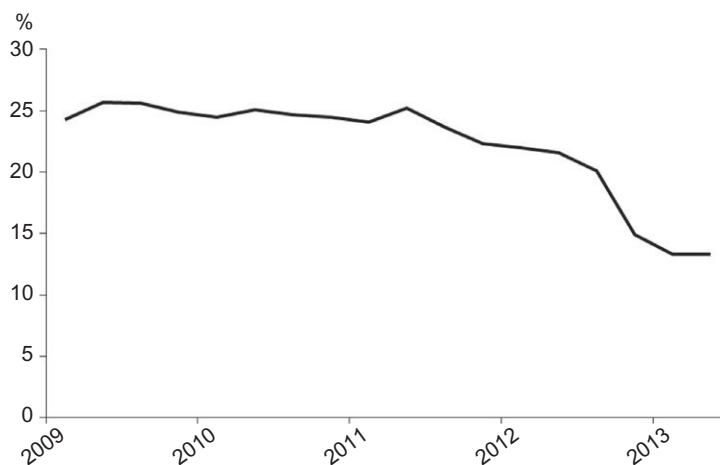


Figure 15.10 Fraction of mortgages with negative equity. *Source: CoreLogic. Fraction of mortgages where the debt outstanding is greater than the estimated property value.*

⁴⁵ For example, [Haughwout et al. \(2013\)](#) reported that on average, households spend \$2500 (in 2012 dollars) in additional home maintenance and improvements during the first year in a house.

source of revenue to local governments is transfer fees levied on property sales.⁴⁶ The turnover rate, therefore, can have implications for local economic activity beyond housing. As noted earlier, housing investment expenditures represent an important offset to the physical depreciation rate for the local housing stock and are an additional source of local economic activity.

15.4.1 Negative equity and strategic default

When a household in negative equity can afford to continue making the mortgage payments, they still have an option to default on the mortgage, in which case the house is sold in foreclosure. We will label this as “strategic default.” If the mortgage is nonrecourse, the cost to the household from making this choice is that their credit is severely damaged for several years, restricting their access to and raising the cost of future borrowing. As we will make clear, to ascribe the house sale to the negative equity, it is important to restrict our attention to cases of strategic default. This is in contrast to a traditional default that results from the combination of an adverse income shock and negative equity. This type of default is driven by the adverse income shock, which requires that the house is sold, whereas the negative equity simply determines how the house is sold—by the owner or by the bank.

Estimating the relative importance of strategic default has been challenged by the lack of data on individual borrower income and mortgage payments. Consequently, early attempts at measuring strategic default used methods of indirect inference. For example, [Experian–Oliver Wyman \(2009\)](#) used credit records to identify potential strategic defaulters. They stipulated that strategic defaulters must meet two conditions: The borrower goes straight from current to default on their mortgage, and the borrower continues to pay their other credit obligations for 6 months after going 60 days delinquent on the mortgage. Using these criteria, they find that 18% of defaults appear to be strategic. A weakness of this approach is that it does not control for two key factors—negative equity and absence of adverse income shocks.

[Bradley et al. \(2012\)](#) were able to remedy this deficiency. The authors use monthly mortgage servicing data that are merged to credit and payroll data. Their sample is selected towards borrowers who work in large firms that are more likely to outsource their employment/income verification to a credit bureau. Using the Experian–Oliver Wyman criteria, they identified 21% of defaults as strategic. Adding in the negative equity and the absence of an adverse income shock reduces this to 14%. The authors reported that the relative prevalence of strategic default is increasing in a borrower’s credit score as well as current LTV. Their results suggest that despite the severe magnitude of negative equity created during the housing bust, only a minority of defaults appear to be strategic in nature.

⁴⁶ See [Lutz et al. \(2011\)](#) for a detailed analysis of the impact of the housing boom and bust on state and local government revenues.

15.4.2 Negative equity and housing turnover

How might negative equity impact the turnover rate of housing? For this discussion, we focus on the turnover of owner-occupied (as opposed to rental) housing. In this context, a house turns over when it is sold and ownership is transferred. Prior to the housing bust, the incidence of negative equity was relatively low. As a consequence, the early literature focused on house price declines that reduced the borrower's equity, but not to the degree that the household was in negative equity. Attention was focused on situations where an owner is both selling a home and buying a subsequent home. [Stein \(1995\)](#) argued that equity reductions resulting from house price declines can limit mobility since the borrower will not have sufficient remaining equity to fund the down payment on a subsequent purchase of a similar or higher-priced home. That is, to move to another ownership situation, the borrower would have to either make up part of the new down payment from other financial assets or move to a less expensive home.⁴⁷ In either case, price declines create a financial friction to selling.

The financial friction to selling created by declining house prices is magnified if these price declines are sufficient to place the household in negative equity. To sell the home, a household in negative equity must be able to pay off the balance of the mortgage.⁴⁸ This requires the household to have the financial assets to be able to make up the deficiency between the balance on the mortgage and the proceeds from selling the home less the transactions costs, creating a financial friction to selling regardless of the household's subsequent tenure decision. Furthermore, to both sell and buy a subsequent home, the household faces an even higher hurdle in that it must pay off the deficiency as well as completely fund a new down payment.

The overall predicted impact of negative equity on owner-occupied housing turnover depends on the relative importance of strategic default. If strategic default is uncommon, then *holding constant the borrower's income*, we would expect that negative equity would reduce housing turnover. That is, the financial friction to selling a home without a default would outweigh the strategic default effect on overall housing turnover. However, if strategic default accounts for a significant amount of overall default, then negative equity could increase housing turnover.⁴⁹ As discussed above, estimates suggest

⁴⁷ Additionally, a household could move from ownership to rental. Here, the friction is the ability to find a similar home and/or location in the rental market. Default would not be a rational outcome so long as the household has enough remaining equity to cover the transactions costs from selling the home.

⁴⁸ The lender could negotiate a short-sale agreement with the household where the household sells the home and the lender agrees to accept less than the outstanding balance on the mortgage. Short sales represented only 6.6% of total home forfeitures in the first quarter of 2008 and rose to 25.2% of home forfeitures by the fourth quarter of 2013. See [OCC Mortgage Metrics Report](#), various issues.

⁴⁹ If the incidence of strategic default is higher the larger the degree of the negative equity, then negative equity may have a nonlinear impact on housing turnover. Low levels of negative equity may reduce turnover, while high levels of negative may increase turnover. The possibility of a U-shaped relationship between negative equity and housing turnover is discussed in [Andersson and Mayock \(2013\)](#).

that most defaults are not strategic. A consequence is that a vast majority of housing turnover related to mortgage default is likely the result of adverse income shocks and not directly due to negative equity.

Empirical work on the relationship between negative equity and mobility divides importantly along the definition of what constitutes a “move.” We will focus our discussion on studies using household level data.⁵⁰ Chan (2001), Engelhardt (2003), Ferreira et al. (2010, 2011), and Andersson and Mayock (2013) focused on moves that involve a transfer of ownership. In contrast, Schulhofer-Wohl (2011), Bucks and Bricker (2013), and Coulson and Grieco (2013) focused on changes of household residence, regardless of whether the ownership of the property has changed hands. The distinction is important since not all changes of household residence involve a change in ownership, whereas changes in ownership nearly always involve as well a change in household residence.

The appropriate definition of a move depends on the specific questions being investigated. For studying the impact of negative equity on local housing markets, we argue that housing turnover is the relevant concept—that is, moves that involve the sale of the home. In contrast, for studying the impact of negative equity on labor markets, household turnover is the more relevant concept. Job mobility is an important element of the functioning of a labor market. Job changes can occur with or without housing turnover. Given our focus on local housing markets, we will focus on moves that entail the sale of a home.⁵¹

For moves that involve the transfer of ownership in the home, the evidence supports the hypothesis that negative equity limits mobility. However, no study provides a clean test because either changes in ownership are not reliably identified or strategic defaults are not differentiated from traditional defaults in the analysis. Chan (2001) provided the earliest evidence using mortgage servicing data from Chemical Bank on ARMs originated between November 1989 and January 1994 in NY, NJ, and CT. She observes if and when a mortgage prepays. A prepayment can result from a borrower either selling the home and paying off the mortgage or refinancing the mortgage. While Chan cannot distinguish between home sales and refinances in her full data, she showed that for a subset of the ARMs originated in New York City where she merged in deeds records information, few prepayments of ARMs over this time period represent refinances. Mortgage defaults are treated as censored at the date of the default. The home’s appraised value at origination is updated using county-level repeat-sales indices to create an updated LTV for each mortgage and demographic variables from the mortgage application (including the borrower’s marital status, number of children, age, education, and whether the borrower is a first-time homeowner) are used to control for many other determinants of household

⁵⁰ For examples of aggregate studies, see Donovan and Schnure (2011) and Nenov (2010).

⁵¹ See Cunningham and Reed (2013) and Valletta (2013) for examples of analysis of negative equity and labor markets.

mobility. Using a proportional hazard framework, Chan found that borrowers with current LTVs above 0.95 are 24% less likely to prepay (therefore move) over a 3-year period as compared to a similar borrower with a current LTV below 0.95.

Engelhardt (2003) used data from the National Longitudinal Survey of Youth (NLSY79) for the period 1985–1996. This has the advantage of being a national survey. He used address information from the confidential survey data to create his mobility measure. His intention is to restrict his mobility measure to moves involving owner to owner transitions and he imposed this restriction using a question in the NLSY79 on homeownership. His resulting mobility measure, therefore, is likely an imperfect indicator of whether a household moves and ownership is transferred. First, cases where the house is sold and the household transitions to renting are censored in the analysis. Second, in cases where the household moves but chooses to rent rather than sell the house, the NLSY79 does not identify the ownership status of the prior home. As such, Engelhardt had to censor these transitions as well. He attempted to distinguish between reductions in mobility due to declines in equity and to loss aversion. He controlled for the origination rather than the current LTV, the extent of any nominal loss that has occurred since the house was purchased, and interaction effects between the origination LTV and loss variables. Since the nominal loss is based on self-reported house values, which may be subject to measurement error, he instrumented the nominal loss variables using implied losses based on repeat-sales house price indices. The IV strategy results in a significant increase in the standard errors of the nominal loss variables making inference more difficult. Engelhardt interpreted his findings as relatively more supportive of loss aversion rather than equity effects as the primary source of the mobility friction from declining house prices.

A challenge for these earlier studies is that they predate the housing bust and so have limited amounts of negative equity in their data. Ferreira et al. (2012) provided the first analysis that captures the initial effects of the housing bust. The authors use AHS data from 1985 to 2009.⁵² The AHS is a biannual survey over this period. Even including house price declines that occurred up to 2007, the reported incidence of negative equity is still only 3.7%. The AHS provides a self-reported house value. Ferreira et al. (2012) acknowledged the likely measurement error in the self-reported house values and instrumented this variable using a house value created from the appraised value updated using house price indices.

The AHS data are well suited for analyzing mobility in that it contains a wealth of demographic information about the household that can affect mobility. A challenge in the AHS data, though, is inferring changes in ownership of a home. The AHS is a panel of housing units, not households. The survey asks if a home is owned or rented, and for a home that is owned, changes in ownership can be observed if it is sold and a new

⁵² This is an update to their earlier study Ferreira et al. (2010) that used AHS data from 1985 to 2007.

household moves in. The “year purchase” variable confirms that the ownership has changed. However, if a home transitions from owned to rented, it is not immediately possible to know if the home has been sold since the year purchased variable is not reported in the case of a rental. If the rented home subsequently transitions back to owned, then it is possible to resolve whether and when the home was sold. As described in [Ferreira et al. \(2012\)](#), in a significant number of cases of owned to rental transitions, the original owners of the home are observed returning at a later date, implying that it was not sold. Consistent with this finding, [Bucks and Bricker \(2013\)](#) reported using panel data from the Survey of Consumer Finances (SCF) for 2007 and 2009 that roughly one-third of homeowners that moved between 2007 and 2009 did not sell the home they owned in 2007. If a different household is observed when the home reverts back to owned, then we can infer that it was sold, and the year purchased variable identifies the timing of the sale. However, if the home is still rented as of the last available survey, it is impossible to ascertain whether a sale took place or not. [Ferreira et al. \(2010, 2012\)](#) decided to censor these transitions.

Another limitation of the AHS data is that for a household that moves, it is not possible to know if the household either suffered an income loss or defaulted on the mortgage in the 2 years between surveys.⁵³ As such, moves associated with a traditional or a strategic default cannot be identified. An advantage, though, of their analysis is that all moves that are identified involve a transfer of ownership. However, for negative equity households, their move definition suffers from two forms of misclassification. First, some of the identified moves occurred as the result of a traditional default where the home is purchased out of foreclosure to be used again as an owned residence. These moves are not the result of the negative equity but will be reflected in the negative equity coefficient since adverse income shocks are left out of the specification.⁵⁴ Second, if following a strategic default a home is purchased out of a foreclosure to be used as a rental property (with no subsequent observed transition back to owned status), then this change in ownership is censored when it should be treated as a move resulting from the negative equity.⁵⁵

With these caveats in mind, [Ferreira et al. \(2010, 2012\)](#) found that negative equity is associated with around a 30% decline in mobility. Given the small fraction of households identified in negative equity, they do not test for variations in the degree of negative equity. In addition, the authors test for mobility effects from financial frictions arising from FRMs in a rising interest rate environment. Since most FRMs are neither assumable nor portable, as mortgage rates rise, a household would have to pay a higher annual

⁵³ [Ferreira et al. \(2010, 2012\)](#) controlled for household income at the beginning of the 2-year period used to track any move as well as the change in household income over the prior 2-year period.

⁵⁴ If controlling for the other explanatory variables in the model the likelihood that a household suffers an adverse income shock is higher in areas with greater declines in house prices, then this will tend to create a positive bias to the negative equity coefficient.

⁵⁵ However, in cases of strategic default where the home is owner-occupied following the foreclosure, the move will be correctly recorded.

mortgage payment to move and take out an equivalent size new mortgage. Ferreira et al. found that \$1000 of additional annual mortgage payments reduces mobility by 16%. They cross validated this finding by comparing it to the impact of the financial friction arising from California's Proposition 13 limitation on property taxes. They reported that \$1000 of annual property tax subsidy to a homeowner reduces mobility by 10%.

Andersson and Mayock (2013) provided the cleanest measures of negative equity as well as moves that involve a sale of the home. A limitation of their study is that it covers only properties in Florida. However, their data span the period from 1999 to 2011, and given the significant declines in Florida house prices in the bust, they have the largest sample of negative equity households. Unlike the AHS and NLSY79 surveys that rely on self-reported house values, they have annual assessed values for each property as of January of each year. To match the timing of their other data, they updated these values to June using county-level repeat-sales price indices. Their mobility variable is derived from deeds records data provided by DataQuick, allowing them to identify all home sales.⁵⁶ They merged this to annual credit bureau data (for June) that allows them to observe the remaining balance and delinquency status of each mortgage. This allows the authors to create a relatively clean updated LTV. They also distinguished between home sales that are associated with a mortgage delinquency or not. However, they did not attempt to distinguish between traditional and strategic defaults. They classified a home sale without a default as a "voluntary" sale and a home sale with a default as an "involuntary" sale. This classification involves potential measurement error in that strategic defaults are categorized as involuntary rather than voluntary.

The steep declines in Florida house prices also allow the authors to examine both the effects of the incidence and magnitude of negative equity on mobility. They reported that 19% of their households are in negative equity, with 10% having an updated LTV of 1.3 or higher. They estimated logit and multinomial logit models where they controlled for the borrower's origination credit score, age, and years in the home. They also included time-fixed effects. Negative equity is found to monotonically reduce the likelihood of voluntary moves, and significant negative equity (LTV in excess of 1.3) to increase the likelihood of involuntary moves.⁵⁷ Overall, they estimated that mobility declined by roughly 25% due to reductions in home equity.

As indicated earlier, the definition of a move is important for the estimated relationship between negative equity and mobility. Schulhofer-Wohl (2011), Bucks and Bricker (2013), and Coulson and Grieco (2013) focused on mobility defined as changes in household residence regardless of whether the home is sold. This definition of a

⁵⁶ A sale is identified regardless of whether the home is subsequently used for owned or rental housing. In addition, if a household moves out but does not sell the home, this is correctly identified as no sale.

⁵⁷ Molloy and Shan (2013) examined panel data on household credit files and found that roughly 50% of households move within 2 years of the start of a foreclosure. Only a small fraction of households appear to move back in with their parents.

move is more appropriate for investigating the potential impact of negative equity on the labor market than its effect on the housing market. Schulhofer-Wohl used the AHS estimation sample from [Ferreira et al. \(2010\)](#) and recoded the censored transitions from owned to rental as a move. As [Ferreira et al. \(2012\)](#) demonstrated, this definition includes many temporary moves where a household leaves, rents out the house, and returns at a later date. Refocusing the mobility definition to the household and not the home, Schulhofer-Wohl reported that negative equity is in fact positively related to mobility.

This finding is corroborated in [Coulson and Grieco \(2013\)](#) who used data from the PSID covering the period from 1999 to 2009. Mobility again is defined as a change of residence by a household. Like the AHS, the PSID allows Coulson and Grieco to control for several demographic variables that might be expected to impact mobility. Similar to the AHS data, default is not controlled for in the analysis. Coulson and Grieco reported that their mobility measure increases with the degree that the household is in negative equity. Given the evidence in [Chan \(2001\)](#) and [Andersson and Mayock \(2013\)](#), the contrast in results between the two mobility definitions is likely driven by default related moves, as well as temporary moves resulting from adverse income shocks. Consistent with this, [Bucks and Bricker \(2013\)](#) using SCF data reported that nearly half of homeowners with negative equity that experienced an unemployment spell moved between 2007 and 2009. Future work will have to disentangle the effects of strategic from traditional defaults to better isolate the role of negative equity as opposed to adverse income shocks on housing turnover.

The financial crisis resulting from the housing bust led to aggressive official responses in terms of fiscal and monetary policy. As a consequence, mortgage rates reached very low levels with 30-year FRMs being offered as low as 3.35% in November/December 2013. For borrowers with FRMs that still had sufficient equity in their house, this steep decline in mortgage rates created a strong incentive to refinance their mortgages. However, for borrowers with agency mortgages guaranteed by Freddie Mac or Fannie Mae, many no longer had sufficient equity to fund the required down payment on a refinance. In response, the Home Affordable Refinance Program (HARP) was introduced in March 2009 to allow these borrowers to refinance so long as they had a clean payment history and had originated their mortgage prior to June 2009. In December 2011, changes were made to the HARP program to expand its eligibility and reduce the associated fees. As of December 2013, 3.1 million mortgages have refinanced under HARP.⁵⁸ Borrowers with FHA mortgages can use a streamline refinance program that does not depend on the borrower's current LTV. Since 2009, 1.5 million high LTV FHA borrowers have refinanced using this program.⁵⁹ Combined, a total of 4.6 million high LTV borrowers have refinanced to a lower interest rate. As of January 2014, prime

⁵⁸ See [FHFA \(2013\)](#).

⁵⁹ See [HUD \(2013a\)](#).

conforming 30-year FRMs with coupon rates of roughly 4.5% or less had aggregate balances of \$2.3 trillion.

The HARP and FHA streamline refinance programs have been successful in helping to circumvent the friction to refinancing caused by the steep decline in house prices. Through refinancing, these impacted borrowers were able to significantly lower their required monthly mortgage payments.⁶⁰ This is clearly beneficial to these households. It is important to note, however, that the borrowers with agency mortgages who did a normal refinance along with those refinancing under the HARP program received a nonassumable and nonportable FRM. An implication is that as the economy improves and mortgage rates normalize, these borrowers will face a large financial friction to moving due to their below market rate mortgages. This interest rate friction was documented by Quigley (1987) and updated by Ferreira et al. (2010, 2011).⁶¹ In contrast, since FHA mortgages are assumable, their streamline refinance program should not generate this interest rate lock-in effect.

15.4.3 Negative equity and housing investment

In addition to affecting housing turnover, negative equity may also reduce housing investment. Housing investments improve the flow of housing services, and for the duration of time that the household expects to live in the home, the household directly receives the benefits from the investments. In fact, if the horizon of the household's expected stay exceeds the useful life of the investment, then the household can expect to receive the full value of the investment. In many cases, though, due to the durable nature of many investments, the service flow from the investment may be expected to continue beyond the household's remaining tenure. In these cases, if there was no way for the household to capture the value of the remaining service flow when it sells the home, then this would reduce the incentives for the household to make long-lived investments.

Capitalization of housing investments into the house price acts to extend the horizon of the household when it is considering housing investment decisions. This helps to promote an efficient investment profile by the household regardless of its expected tenure in the home.⁶² However, for capitalization to support efficient investment, at the time of the investment decision, the homeowner must expect to have a positive equity position when the home is to be sold. That is, the homeowner needs to be fully exposed to any gains or losses in the value of the home that are associated with the investment

⁶⁰ Borrowers refinancing under HARP on average lower their monthly mortgage payment by \$137; see Zhu (2012).

⁶¹ Assuming that these borrowers would have to give up the \$137 per month in lower monthly mortgage payments in order to move, this would imply an annual financial friction of \$1644. Using the estimates from Ferreira et al. (2010, 2012), this would be expected to reduce mobility by 16%.

⁶² See Fischel (2001).

decision.⁶³ In normal housing markets, this is not a binding constraint since the homeowner's initial equity tends to rise over time due to a combination of house price increases and debt amortization.

Negative equity may reduce housing investment for a variety of reasons. First, negative equity puts the borrower at risk of a default in the future if the borrower suffers an adverse income shock. In the event of a default, any capitalization from a housing investment will first go to the lender. As a result, negative equity creates an agency problem between the lender and the household. The lender would like the household to continue to make costly housing investments, but the lender has no direct control over these investment decisions until they take title to the property through foreclosure. Second, for those types of investments that can be delayed with little impact on the immediate flow of housing services, the household may choose to postpone these investments in order to create precautionary savings (see [Haughwout et al., 2013](#)). Third, households may need to use the equity in their home to finance larger investment projects, and underwater households do not have access to this source of collateralized financing. Finally, households in negative equity have experienced a wealth loss that can lead to reduced consumption in general and housing investments in particular. We focus here specifically on the negative equity instead of the wealth declines since the negative equity results from a feature of the mortgage contract—no maintenance margin—whereas the wealth effects are less dependent on the structure of the mortgage.⁶⁴

Prior to the housing bust, given the paucity of negative equity, research focused on the behavior of maintenance and improvement expenditures (hereafter housing investment) to adverse income shocks instead of adverse equity shocks.⁶⁵ For most categories of housing investment, the investment can be deferred without a first-order reduction in the flow of housing services. This suggests that the timing of housing investment could be used to help smooth consumption in the face of transitory income shocks. [Gyourko and Tracy \(2006\)](#) used AHS data from 1985 to 1993 to estimate the response of housing investment to transitory income shocks. Using the panel structure of the AHS data, they estimated a heterogeneous growth model for household income. This allows them to estimate the transitory component of the income residual. While they found a statistically significant response of housing investment to the estimated transitory

⁶³ An important question is whether making mortgage loans recourse mitigates the need for the borrower to expect to have a positive equity stake at the time of sale. We will return to this question.

⁶⁴ The potential wealth effect is influenced by one feature of the mortgage contract—whether the lender has recourse or not. In the case of a nonrecourse mortgage, the potential negative wealth effect is limited to the borrower's down payment. In the case of a recourse mortgage, the lender can petition to pursue the borrower's other financial assets to satisfy any deficiency.

⁶⁵ An earlier literature related property tax delinquencies, which in the extreme can generate negative equity, to housing disinvestment and ultimately abandonment. See [White \(1986\)](#), [O'Flaherty \(1993\)](#), and [Scafidi et al. \(1998\)](#).

income shocks, this component plays a relatively minor overall role in the household's overall strategy for smoothing its consumption.⁶⁶

Turning to negative equity effects on housing investment, [Haughwout et al. \(2013\)](#) and [Melzer \(2012\)](#) both used the CEX data to investigate this relationship.⁶⁷ The CEX consists of short panels—essentially four consecutive quarterly surveys—of households. An advantage of the CEX is that it has detailed information on a wide range of types of housing investment as well as detailed demographic information on the household. Given the lumpy nature of most housing investments, Haughwout et al. aggregated these expenditures across the four surveys, whereas Melzer estimated his specification using the quarterly data. Negative equity must be inferred from the household's self-reported house value and estimates of the current outstanding balances for all loans secured by the house. Haughwout et al. created a negative equity indicator based on the initial quarter survey information and then instrumented this using negative equity indicators based on the remaining three quarters. They argued that an important component of the measurement error in self-reported house values is transitory. Melzer did not try to address measurement error. Another specification difference is that Haughwout et al. controlled for household income, whereas Melzer controlled for total household expenditures. The IV results from Haughwout et al. indicate that negative equity is associated with a 74% (\$2610) annual reduction in housing investment. Melzer reported a much smaller impact of 30% or \$800 per year. The income elasticity reported by Haughwout et al. of 0.58 exceeds the elasticity of 0.42 reported in [Gyourko and Tracy \(2006\)](#).

An interesting question is whether we can identify the channels that may be generating this estimated effect for negative equity on housing investment. Earlier, we discussed that these channels include the agency problem regarding the investment decision, a demand for precautionary savings, collateral constraints on borrowing, and wealth effects. The precautionary savings, collateral, and wealth channels have the feature that they would also be expected to impact other significant durable purchases such as home furnishings and vehicles. However, since these categories of durable goods are not tied to the home, they should not be impacted by agency problems arising from negative equity.⁶⁸ Melzer reported no significant impacts of negative equity on either vehicle purchases or home furnishings and equipment expenditures. These results provide

⁶⁶ They also found that housing investment is positively related to estimated permanent income shocks. This can be interpreted as a wealth effect on housing investment.

⁶⁷ Haughwout et al. used data from 2007 to 2012, while Melzer used data from 2006 to 2011.

⁶⁸ There are some measurement issues with these two variables. The home furnishings category include large appliances such as refrigerators that are installed but not as part of a contracted project. Similar installed appliances that are part of a contracted project are treated as home improvement expenditures. This creates an asymmetry in the treatment of these appliances. Also, the category vehicle purchases includes both purchases and leases. For leases it is not possible to distinguish between a new lease and an existing lease. This makes it impossible to isolate new spending decisions on vehicles.

indirect support for the agency channel being important for the estimated negative equity and housing investment relationship. However, the other channels through which declining house prices may impact durable goods purchases would still be expected to lead to a reduction in expenditures in these categories.

There are two ways in which lenders can protect themselves from this agency problem regarding housing investments. The first is to require larger down payments. This would reduce the likelihood that the borrower ends up in negative equity. The second is to attempt to make the borrower internalize the consequence of a decision not to maintain the property even after the borrower no longer has any home equity. A potential legal remedy that attempts to do this is making the mortgage a recourse loan. Recourse mortgages allow the lender to file a deficiency judgment against the borrower for any shortfall between the remaining mortgage balance and the proceeds from selling the home in foreclosure. The lender may satisfy the deficiency judgment using the borrower's other financial assets.

In practice, recourse is unlikely to discipline borrower housing investment decisions in most circumstances. For recourse to provide an effective legal remedy to the agency problems created by negative equity, three things need to hold. First, the borrower must be aware that the mortgage is a recourse loan and understand the potential financial consequences. If borrowers are not aware of this legal right by lenders, then recourse will not affect their housing investment decisions. [Haughwout et al. \(2013\)](#) reported that among US borrowers with recourse mortgages, only 56% correctly identified the recourse status of the mortgage.⁶⁹ Second, at the time that the borrower is considering making a costly maintenance decision, the borrower must expect to have financial assets that will exceed the amount of negative equity at the time of the foreclosure completion. If the negative equity already exceeds the borrower's financial assets, then any further decline in the value of the home due to a decision not to carry out a specific maintenance project will not have any impact on the expected collection from a deficiency judgment. [Haughwout et al. \(2013\)](#) reported that for borrowers with at least \$10,000 in negative equity in their CEX estimation sample, only 22% had current financial assets that exceeded \$10,000. Finally, filing for a deficiency judgment is costly with the specific expense varying by state. Lenders will only file if the expected claim exceeds the filing costs. This suggests that lenders will only likely file in cases with relatively large deficiency amounts. The [FHFA's Inspector General Report \(2012\)](#) indicated that the GSEs filed for deficiency judgments in 2011 for only 10.3% of all foreclosures involving mortgages guaranteed by the GSEs.

The impact of negative equity on housing investment may help to explain the findings of negative price externalities for foreclosures on nearby properties. Much of this literature focuses on the flows into foreclosure, which we discuss in [Section 15.5](#).

⁶⁹ This is from the 2013 Survey of Consumer Expectations.

However, [Gerardi et al. \(2012\)](#) expanded the focus to look at the stock of seriously delinquent properties. They found that the negative price externality is modest and emerges with serious delinquency, peaks during foreclosure, and dissipates within a year following the sale by the lender. An explanation for the magnitude and timing of the effect may be the reduced investment by underwater borrowers. Unfortunately, the CEX data do not have information on borrower delinquency, so [Melzer \(2012\)](#) and [Haughwout et al. \(2013\)](#) cannot determine if the disincentive to make costly housing investments has its onset when a borrower enters negative equity or only later when the borrower reaches some level of delinquency. In addition, the finding by [Gerardi et al. \(2012\)](#) that the negative externality dissipates within a year following a sale by the lender could be explained if most of the foregone maintenance is made up by the new owner.⁷⁰ This possibility points to a fruitful avenue for future research using the AHS where one can observe the maintenance decisions by the subsequent owner.

Before turning to the issue of foreclosures, it is worth discussing how the structure of mortgage contracts may impact the social benefits from homeownership. A wide range of government policies have been directed towards increasing the homeownership rate on the belief that homeowners take better care not only of their properties but also of their communities.⁷¹ [Fischel \(2001\)](#) argued that homeowners make these costly investments because they believe that the investments will enhance the value of their homes. This capitalization effect, though, assumes that the homeowner has positive equity. [Haughwout et al. \(2010\)](#) showed that following the housing bust, the median owner in several metro areas located in the boom/bust states had negative equity. In addition to a diminished incentive to invest in their homes, pervasive negative equity could lead to underinvestment in local public infrastructure. This would be another area for future research.

15.5. FORECLOSURES

[Figure 15.11](#) shows the dramatic increase in foreclosures following the housing market bust. The foreclosure rate rose rapidly from 0.6% of mortgages at the start of 2007 to 3.0% in 2010, peaking at 3.9% in 2012.

15.5.1 The distribution of foreclosures

There are many reasons to expect a pronounced spatial pattern in foreclosures. As mentioned earlier, many studies provide evidence that subprime mortgage lending was disproportionately made in nonwhite and poorer neighborhoods. These loans, of course,

⁷⁰ Alternatively, the lender may remedy some of the maintenance issues prior to the sale.

⁷¹ Homeowners tend to vote for local school bond issues ([Bergstrom et al., 1982](#) and [Hilber and Mayer, 2009](#)). Homeowners are also more likely to participate in civic activities such as voting in local elections ([DiPasquale and Glaeser, 1999](#)).

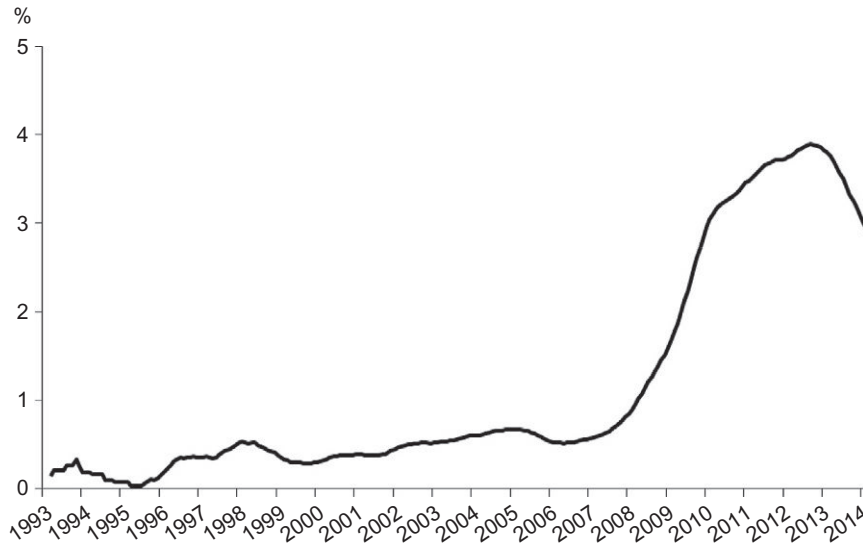


Figure 15.11 Fraction of mortgaged homes in foreclosure. Source: Lender Processing Services. First-lien, owner-occupied purchase mortgages.

turned out to be much quicker to default during the bust phase of the housing cycle.⁷² Figure 15.12 provides an illustration of the geographic concentration of default in predominantly black and Hispanic neighborhoods in New York City. As in Figure 15.8, the shading shows the share of black or Hispanic residents in each census tract, but here, each dot represents one residential foreclosure filing issued in 2009. As expected, Figures 15.8 and 15.12 look remarkably similar. Haughwout et al. (2012) also found a spatial pattern in the prevalence of investor-originated loans, which ex post turned out to be quick to default when house prices began to fall.

House price cycles were more pronounced in some places, leading to a higher probability of negative equity. Income shocks were also spatially concentrated, in many cases in the same areas as the biggest house price cycles. These phenomena varied significantly both across and within metropolitan areas. An important line of inquiry is the extent to which the high foreclosure rates for certain vintages of nonprime mortgages were due to weaker underwriting of these mortgages—in both observable risk characteristics and unobservable dimensions—and how much to the deteriorating economic environment. Haughwout et al. (2008) found that observed factors explain less than half of the sharp rise in early subprime defaults between 2003 and 2007. Around 70% of the explained variation is due to economic factors, not to borrower or loan risk characteristics. The authors speculate that nondeclared investors might explain the large unexplained component of

⁷² For example, see Chan et al. (2013) and Bayer et al. (2014).



Figure 15.12 Distribution of foreclosure filings in NYC, by census tract race and ethnicity. *Source: US Census 2000 and Public Data Corporation 2009. Each dot represents one residential foreclosure filing (lis pendens). Shading shows the fraction of non-Hispanic blacks and the fraction of Hispanics of any race among each tract's population.*

early defaults. [Palmer \(2014\)](#) extends the analysis of [Haughwout et al. \(2008\)](#) by looking at default behavior over several years following origination. Palmer finds that origination year vintage effects capturing any unobserved underwriting quality explain little of the observed differences in default behavior.

Property tax lien foreclosures have also become an increasing problem in recent years. While distinct from mortgage foreclosures, they share some of the same underlying causes and have similar external impacts on the surrounding neighborhood. A local government can generally place a tax lien on the property if a homeowner is delinquent for more than 1 year on property taxes or other municipal charges. These tax liens have priority over any mortgage and give the lien holder the right to foreclose in order to recover the deficiency.

Failure to pay property taxes has been discussed earlier in the context of reverse mortgages. For forward (i.e., nonreverse) mortgages, lenders typically require an escrow account that covers property taxes and insurance. However, the majority of subprime loans made prior to 2008 did not include an escrow account, perhaps to make the monthly loan payment seem more affordable ([Rao, 2012](#)). Thus, higher cost subprime mortgage borrowers, who were more prevalent in low-income and nonwhite

communities, not only had higher mortgage default rates but also are also more likely to be at risk of a tax deficiency. Because local governments tend to impose substantial interest and penalties on any deficiencies and have first-lien priority, lenders have an incentive to pay any nonescrowed taxes, so long as the home's expected present value at a mortgage foreclosure auction exceeds the expected present value of the tax payments. Of course, that condition is less likely to be met in neighborhoods suffering large declines in property values, leading to higher rates of tax foreclosure and all the negative spillover effects associated with mortgage foreclosure that we discuss below.⁷³

15.5.2 Foreclosure externalities

An extensive literature has developed around the important question of whether there exists a negative “externality” of a foreclosure property on the prices of nearby houses (Immergluck and Smith, 2006a; Schuetz et al., 2008; Campbell et al., 2011). This question is important for both positive and normative reasons. On the positive side, understanding spillovers can help to explain observed spatial concentrations in price declines and in mortgage defaults. In addition, the existence of significant spillovers of foreclosures onto neighboring property prices could lend support to the case for interventions to prevent foreclosures in the future, ranging from reforms of the mortgage contract itself to ex post interventions intended to reduce foreclosures after a market downturn.

In general, these studies have found such externalities, with the price of nearby houses falling as more properties in the neighborhood are foreclosed upon. For example, Campbell et al. (2011), using sales from Massachusetts over the period 1987–2009 estimated hedonic price regressions, augmented by information on the number of foreclosures within small distances from the unit in question. In so doing, they are able to control for the effect of local demand shocks, which have the potential to lead to a spurious correlation between foreclosures and local price declines. The results indicate that an additional foreclosure within 0.05 miles (88 yards) of a nonforced sale reduces the price that the unit will command by about 1%.

More work is required to completely identify the channel through which these externalities are produced. Several hypotheses seem plausible. A natural candidate is that increased foreclosures shift the supply curve outward, leading to lower prices in the neighborhood. Two recent studies find support for this hypothesis. Hartley (2011) used data from Chicago and found no effect of multifamily foreclosures on the sales price of nearby single-family units. Given the segmentation between demands for those two types of housing, the negative spillover identified in other work must be attributable to a supply effect.

⁷³ The sale of property tax liens to third-party investors, and in some cases, their securitization has led to a host of other problems affecting not only properties with mortgages. See Rao (2012) for more details.

Anenberg and Kung (2013) used national data and exploit the timing of REO (real estate owned—that is, properties owned by banks) listings to estimate the effect of foreclosures on neighboring properties. The authors found that only when the property is actually listed for sale—not before it is listed or after it is sold—do neighboring prices decline. In this case, the authors find a 1.6% reduction in sale prices of nearby homes when the REO is listed, an effect that diminishes with distance from the REO and becomes zero at roughly $\frac{1}{2}$ mile. As noted, the authors find that the price declines are temporary and reverse as soon as the foreclosure property is sold. The authors argue that this overall temporal pattern is consistent with the externality being produced by price competition.

As discussed earlier, a second possible channel to consider is diminished maintenance of the foreclosed structure, creating a localized disamenity that Gerardi et al. (2012) argued is important. The bids of prospective buyers of neighboring properties may be reduced by proximity to an undermaintained unit in the foreclosure process, which creates a neighborhood negative spillover. Interestingly, Anenberg and Kung (2013), while finding that the spillover from foreclosed properties is primarily a supply shift, conducted additional analysis on high-density, low-price neighborhoods and found that in those areas, the data support the existence of a disamenity effect. Foreclosure properties in such neighborhoods are probably in poorer condition to begin with, and they may be especially likely to generate a negative spillover with a further reduction in maintenance.

We do not know much about the timing of this reduction in maintenance, although Melzer (2012) and Haughwout et al. (2013) found a diminution in housing investment from the onset of negative equity, at which point the owner's incentive to maintain becomes less clear-cut. We also do not know how persistent these shortfalls are. Do new owners make them up? The persistence of the neighborhood effects identified in Campbell et al. implies that they may not do so immediately. Another potential channel through which foreclosures may affect neighboring property values is crime. Ellen et al. (2013) and Immergluck and Smith (2006b) found that neighborhoods with additional foreclosures experience increases in crime, which presumably will reduce property values.

A second important effect of the spatial concentration of foreclosures is reduced property tax revenues because the owner of the foreclosed property has little incentive to continue making tax payments on a property that he will not own much longer. In addition, the subsequent reduction in neighboring property values contributes to further reductions in the tax base, leading to diminished services or higher tax rates, each of which will put further downward pressure on home values. This is another mechanism that can generate the negative price externalities found in previous work, although the precise spatial patterns found by previous authors suggest that at least some mechanisms are operating at a geographic scale smaller than the municipality.

Beyond the effects of foreclosure on locations, of course, is the significant effect on individual borrowers who must vacate their homes. These individuals experience large reductions in their credit ratings that take many years to repair (Brevoort and Cooper, 2013), reducing their ability to smooth earnings shocks and invest in their futures. The impact on other household members who are also forced to move may also be substantial. For example, Been et al. (2011) found that New York City elementary and middle school students living in foreclosed buildings were, all else equal, more likely to switch schools (generally regarded by experts as detrimental), and on average, they switched to schools offering academically weaker peers.

The rapidly growing literature on foreclosure externalities has produced some important facts. We now know that foreclosures exert a negative effect on nearby properties—generally those within half a mile or less of the foreclosed property. Those external effects appear to be short-lived and may be primarily attributable to price competition rather than a disamenity effect. This suggests that isolated foreclosures will not have persistent, widespread effects on most neighborhoods. However, it is worth recalling that during the recent bust, foreclosures themselves tended to be spatially concentrated, especially in the poorer, denser neighborhoods that received large amounts of subprime credit. The existing evidence clearly indicates that these properties will be undermaintained, and in these neighborhoods, there exists some evidence that spillovers are more persistent and more related to the disamenity. More research in this area is likely warranted to ensure that we thoroughly understand these important dynamics.

15.6. CONCLUSION

Our review of the effects of mortgage finance on spatial patterns of residential activity indicates that while we understand much of how these effects play out, we still have much to learn. The events of the 2000s provide a wealth of information on which to base future analyses of these important questions. One element that we have not discussed much is whether alternatives to the current form of mortgages could improve outcomes.

Despite many innovations and changes to the system of mortgage finance in the United States, the core product, the 30-year amortizing mortgage that replaced the standard short-term balloon mortgage in the 1930s, has remained essentially unchanged for almost a century. In the interim, American society and the US economy have undergone fundamental changes. Discrimination and differential outcomes by race and ethnicity have become more repugnant to society, leading to a more acute focus on the role of mortgage finance in perpetuating disparities. The increasingly high cost of medical and long-term care, the sustained increases in longevity, and the declining prevalence of traditional defined benefit pensions have highlighted housing's role as an important source of retirement and precautionary saving. The last 20 years has seen dramatic fluctuations in housing prices, and these have had greater impact on households as the ability

to borrow against home equity has expanded. Further, recent decades have seen substantial increases in income inequality that may well be exacerbated by the current system of mortgage finance and by some aspects of housing-related public policies, including the tax code and mortgage industry regulation.

Ideally, an alternate mortgage instrument would alleviate the down payment constraint associated with the transition to homeownership, eliminate any scope for discrimination in mortgage qualification or pricing, allow households to easily tap into accumulated home equity while also serving as an automatic vehicle for saving, reduce the frictions and costs arising from negative equity and foreclosure, and reduce the exposure of households to fluctuations in the housing market. Such an instrument does not yet exist, although several proposals have the potential to address a subset of these issues. The housing partnerships proposed by [Caplin et al. \(1997a\)](#) and described earlier in this chapter can ease the transition to homeownership and reduce housing asset risk and the consequences of negative equity. [Shiller \(2008\)](#) had proposed continuous work out mortgages (CWMs) in which the mortgage balance and payment schedule would systematically and continuously adjust to a local home price index, as well as other economic indices such as the unemployment rate, in a way that preserves some home equity and payment affordability for the borrower. For lenders, CWMs may be attractive as they reduce the risk of foreclosure costs and inadequate home maintenance.

As [Shiller \(2014\)](#) notes, the development of new mortgage forms is costly, while the benefits are difficult to capture privately. Whether these and other mortgage product innovations will grow in importance depends critically on research, as well as on industry and government support for experimentation. The 30-year mortgage that is now taken for granted was the result of government intervention in the mortgage market, and further beneficial innovations will not necessarily arise naturally from the private sector.

ACKNOWLEDGMENTS

Thanks to Shaopeng He, Sean Capperis, Neha Hridaya, and Samuel Kapon for assistance with the figures and bibliography. We thank Karen Pence, Stijn van Nieuwerburgh, and Will Strange for their helpful comments. The views represented here are those of the authors and do not necessarily reflect those of the Federal Reserve Bank of New York or the Federal Reserve System.

REFERENCES

- Anderson, F., Mayock, T., 2013. How does home equity affect mobility? Working paper Office of the Comptroller of the Currency (September).
- Andrews, D., Sanchez, A.C., 2011. The evolution of homeownership rates in selected OECD countries: demographic and public policy influences. *OECD J.* 2011, 207–243.
- Anenberg, E., Kung, E., 2013. Estimates of the Size and Source of Price Declines Due to Nearby Foreclosures. Board of Governors, Washington, DC (December).
- Apgar, W., Duda, M., 2003. The twenty-fifth anniversary of the Community Reinvestment Act: past accomplishments and future regulatory challenges. *Econ. Pol. Rev.* 9, 169–191.

- Avery, R.B., Calem, P.S., Canner, G.B., 2003. The effects of the Community Reinvestment Act on local communities. In: Proceedings 878. Federal Reserve Bank of Chicago, Washington, DC.
- Avery, R.B., Brevoort, K.P., Canner, G.B., 2006. Higher-priced home lending and the 2005 HMDA data. *Fed. Reserv. Bull.* 92, A123–A166.
- Barr, M.S., 2005. Credit where it counts: the Community Reinvestment Act and its critics. *N. Y. Univ. Law Rev.* 80, 513–652.
- Bayer, P., Ferreira, F., Ross, S.L., 2014. Race, age and high cost mortgage lending: borrower experiences leading up to the crisis. Working paper. Duke University.
- Been, V., Ellen, I.G., Schwartz, A.E., Stiefel, L., Weinstein, M., 2011. Does losing your home mean losing your school? Effects of foreclosures on the school mobility of children. *Reg. Sci. Urban Econ.* 41, 407–414.
- Bergstrom, T.C., Rubinfeld, D.L., Shapiro, P., 1982. Micro-based estimates of demand functions for local school expenditures. *Econometrica* 50, 1183–1205.
- Bernanke, B.S., Board of Governors, 2007. The Community Reinvestment Act: its evolution and new challenges. In: Community Affairs Research Conference, Washington, D.C., March 30. <http://www.federalreserve.gov/newsevents/speech/bernanke20070330a.htm>.
- Bocian, D.G., Ernst, K.S., Li, W., 2008. Race, ethnicity and subprime home loan pricing. *J. Econ. Bus.* 60, 110–124.
- Bond, P., Elul, R., Garyn-Tal, S., Musto, D.K., 2013. Does junior inherit? Refinancing and the blocking power of second mortgages. Working paper No. 13-3/R. Federal Reserve Bank of Philadelphia (June). <https://www.philadelphiafed.org/research-and-data/publications/working-papers/2013/wp13-3R.pdf>.
- Bourassa, S.C., Grigsby, W.G., 2000. Income tax concessions for owner-occupied housing. *Hous. Policy Debate* 11 (3), 521–546.
- Bourassa, S.C., Haurin, D.R., Hendershott, P.H., Hoesli, M., 2013. Mortgage interest deductions and homeownership: an international perspective. *J. Real Estate Lit.* 21 (2), 181–203.
- Bradley, M.G., Crews Cutts, A., Gosh, B.K., Liu, W., 2012. Uncovering the risks of strategic default. Working paper. CoreLogic.
- Brevoort, K.P., Cooper, C.R., 2013. Foreclosure's wake: the credit experience of individuals following foreclosure. *Real Estate Econ.* 41 (4), 747–792.
- Bucks, B.K., Bricker, J., 2013. Household mobility over the great recession: evidence from the U.S. 2007–09 survey of consumer finances panel: Finance and Economics Discussion Series 2013-53. Federal Reserve Board.
- Campbell, J.Y., Giglio, S., Pathak, P., 2011. Forced sales and house prices. *Am. Econ. Rev.* 101, 2109–2131.
- Caplin, A., Chan, S., Freeman, C., Tracy, J., 1997a. Housing Partnerships. MIT Press, Cambridge, MA.
- Caplin, A., Freeman, C., Tracy, J., 1997b. Collateral damage: refinancing constraints and regional recessions. *J. Money Credit Bank.* 29, 496–516.
- Chan, S., 2001. Spatial lock-in: do falling house prices constrain residential mobility? *J. Urban Econ.* 49, 567–586.
- Chan, S., Gedal, M., Been, V., Haughwout, A., 2013. The role of neighborhood characteristics in mortgage default risk: evidence from New York City. *J. Hous. Econ.* 22, 100–118.
- Chatterjee, S., Corbae, D., Nakajima, M., Rios-Rull, J.-V., 2007. A quantitative theory of unsecured consumer credit with risk of default. *Econometrica* 75, 1525–1589.
- Consumer Financial Protection Bureau, 2012. Reverse mortgages: report to congress.
- Coulson, N.E., Grieco, P.L.E., 2013. Mobility and mortgages: evidence from the PSID. *Reg. Sci. Urban Econ.* 43, 1–7.
- Courchane, M.J., 2007. The pricing of home mortgage loans for minority borrowers: how much of the APR differential can we explain? *J. Real Estate Res.* 29, 399–439.
- Courchane, M.J., Surette, B., Zorn, P., 2004. Subprime borrowers: mortgage transitions and outcomes. *J. Real Estate Financ. Econ.* 29, 365–392.
- Cox, D., Jappelli, T., 1990. Credit rationing and private transfers: evidence from survey data. *Rev. Econ. Stat.* 72, 445–454.
- Crawford, A., Cesaire, M., Jie, Z., 2013. The residential mortgage market in Canada: a primer. *Financial System Review*, 53–63.
- Cunningham, C., Reed, R.R., 2013. Negative equity and wages. *Reg. Sci. Urban Econ.* 43, 841–849.
- Davidoff, T., 2006. Maintenance and the home equity of the elderly. Working paper. <http://flatbush.sauder.ubc.ca/REerevise.pdf>.

- Davidoff, T., 2010. Home equity commitment and long-term care insurance demand. *J. Public Econ.* 94, 44–49.
- Davidoff, T., 2013a. Can ‘high costs’ justify weak demand for the home equity conversion mortgage? Working paper. Sauder School of Business, University of British Columbia (October). <http://ssrn.com/abstract=2146988>.
- Davidoff, T., 2013b. Supply elasticity the housing cycle of the 2000s. *Real Estate Econ.* 41, 793–813.
- Davidoff, T., 2014. Reverse mortgage demographics and collateral performance. Working paper. (February). <http://ssrn.com/abstract=2399942>.
- DiPasquale, D., Glaeser, E.L., 1999. Incentives and social capital: are homeowners better citizens? *J. Urban Econ.* 45, 354–384.
- Donovan, C., Schnure, C., 2011. Locked in the house: do underwater mortgages reduce labor market mobility? Working paper (May).
- Ellen, I.G., Lacoe, J., Sharygin, C.A., 2013. Do foreclosures cause crime? *J. Urban Econ.* 74, 59–70.
- Engelhardt, G., 1994. House prices and the decision to save for down payments. *J. Urban Econ.* 36, 209–237.
- Engelhardt, G., 2003. Nominal loss aversion, housing equity constraints, and household mobility: evidence from the United States. *J. Urban Econ.* 53, 171–195.
- Engelhardt, G., Mayer, C.J., 1998. Intergenerational transfers, borrowing constraints, and saving behavior: evidence from the housing market. *J. Urban Econ.* 44, 135–157.
- Eriksen, M.D., Greenhalgh-Stanley, N., Engelhardt, G.V., 2013. Home safety, accessibility, and elderly health: evidence from falls. Working paper. Texas Tech University. http://meriksen.ba.ttu.edu/Papers/Falls_Eriksen.pdf.
- Experian–Oliver Wyman Market Intelligence Reports, 2009. Understanding strategic default in mortgages part I. 2009 Topical report series (August). http://www.hofinet.org/upload_docs/Experian-OliverWyman2009OW_Strategic_Mortgage_Default_Study.pdf.
- Fairlie, R.W., Krashinsky, H.A., 2012. Liquidity constraints, household wealth, and entrepreneurship revisited. *Rev. Income Wealth* 58, 279–306.
- Fannie Mae, April 2013. National Housing Survey Q1–2013 Data Summary.
- Ferreira, F., Gyourko, J., Tracy, J., 2010. Housing busts and household mobility. *J. Urban Econ.* 68, 34–45.
- Ferreira, F., Gyourko, J., Tracy, J., 2012. Housing busts and household mobility: an update. *Econ. Pol. Rev.* 18, 1–15.
- FHFA, 2012. FHFA’S oversight of the enterprises’ efforts to recover losses from foreclosure sales. Audit report: AUD–2013–001. Federal Housing Finance Agency, Office of Inspector General.
- FHFA, 2013. Refinance report: fourth quarter 2013. Federal Housing Finance Agency, Washington, DC.
- Fischel, W.A., 2001. *The Homevoter Hypothesis: How Home Values Influence Local Government Taxation, School Finance and Land-Use Policies*. Harvard University Press, Cambridge.
- Genesove, D., Han, L., 2013. A spatial look at housing boom and bust cycles. In: Glaeser, E., Sinai, T. (Eds.), *Housing and the Financial Crisis*. National Bureau of Economic Research, Chicago, pp. 105–142.
- Genesove, D., Mayer, C.J., 1997. Equity and time to sale in the real estate market. *Am. Econ. Rev.* 87, 255–269.
- Gerardi, K.S., Foote, C.L., Willen, P., 2010. Reasonable people did disagree: optimism and pessimism about the U.S. Housing Market before the crash. Public Policy Discussion Paper No. 10–5. Federal Reserve Bank of Boston (September). <https://www.bostonfed.org/economic/ppdp/2010/ppdp1005.pdf>.
- Gerardi, K.S., Rosenblatt, E., Willen, P.S., Yao, V.W., 2012. Foreclosure externalities: some new evidence. Working paper No. 12–5. Federal Reserve Bank of Boston (September). <http://www.nber.org/papers/w18353>.
- Ghent, A.C., Kudlyak, M., 2011. Recourse and residential mortgage default: evidence from U.S. States. *Rev. Financ. Stud.* 24, 3139–3186.
- Ghent, A.C., Hernandez-Murillo, R., Owyang, M.T., 2014. Differences in subprime loan pricing across races and neighborhoods. Working paper 2011–033C. Federal Reserve Bank of St Louis (March). <http://research.stlouisfed.org/wp/2011/2011-033.pdf>.
- Gibb, K., Whitehead, C., 2007. Towards the more effective use of housing finance and subsidy. *Hous. Stud.* 22, 183–200.
- Glaeser, E., Gyourko, J., 2005. Urban decline and durable housing. *J. Polit. Econ.* 113 (2), 345–375.

- Glaeser, E.L., Gyourko, J., Saiz, A., 2008. Housing supply and housing bubbles. *J. Urban Econ.* 64, 198–217.
- Glaeser, E.L., Gottlieb, J.D., Gyourko, J., 2013. Can cheap credit explain the housing boom? In: Glaeser, E., Sinai, T. (Eds.), *Housing and the Financial Crisis*. National Bureau of Economic Research, Chicago, pp. 301–360.
- Goodman, A.C., 2013. Is there an S in urban housing supply? Or what on earth happened in Detroit? *J. Hous. Econ.* 22, 179–191.
- Green, R.K., Wachter, S.M., 2005. The American mortgage in historical and international context. *J. Econ. Perspect.* 19, 93–114.
- Greenspan, A., Kennedy, J., 2008. Sources and uses of equity extracted from homes. *Oxf. Rev. Econ. Policy* 24, 120–144.
- Gyourko, J., Molloy, R., 2014. Regulation and housing supply. Working paper. Wharton School (July).
- Gyourko, J., Sinai, T., 2004. The (un)changing geographical distribution of housing tax benefits: 1980–2000. In: James, P. (Ed.), *Tax Policy and the Economy*, vol. 18. MIT Press, Boston, MA, pp. 175–208.
- Gyourko, J., Tracy, J., 2006. Using home maintenance and repairs to smooth variable earnings. *Rev. Econ. Stat.* 88, 736–747.
- Hanson, A., Hawley, Z., Martin, H., Liu, B., 2013. Experimental tests for discrimination by mortgage loan originators. Working paper. Marquette University. https://editorialexpress.com/cgi-bin/conference/download.cgi?db_name=UEA2013&paper_id=93.
- Harding, J., Rosenthal, S.S., 2013. Homeowner-entrepreneurs, housing capital gains, and self-employment. Working paper. University of Connecticut (September). <https://www.google.com/#q=homeowner-entrepreneurs%2C+housing+capital+gains%2C+and+self-employment>.
- Harding, J.C., Sirmans, C.F., Rosenthal, S.S., 2007. Depreciation of housing capital, maintenance, and house price inflation. *J. Urban Econ.* 61, 193–217.
- Hartley, D., 2011. The effect of foreclosures on nearby housing prices: supply or disamenity? Working paper 10-11R. Federal Reserve Bank of Cleveland (May). https://www.clevelandfed.org/Community_Development/events/PS2010/presentations_papers/Hartley_paper.pdf.
- Haughwout, A., Peach, R., Tracy, J., 2008. Juvenile delinquent mortgages: bad credit or bad economy. *J. Urban Econ.* 64, 246–257.
- Haughwout, A., Mayer, C., Tracy, J., 2009. Subprime mortgage pricing: the impact of race, ethnicity, and gender on the cost of borrowing. In: Burtless, G., Rothenberg Pack, J. (Eds.), *Brookings-Wharton Papers on Urban Affairs*. Brookings, Washington, DC, pp. 33–63.
- Haughwout, A., Peach, R., Tracy, J., 2010. The homeownership gap. *Curr. Issues Econ. Fin.* 16, 1–11.
- Haughwout, A., Lee, D., Tracy, J., van der Klaauw, W., 2012. Real estate investors, the leverage cycle, and the housing market crisis. Staff report No. 514. Federal Reserve Bank of New York (September). http://www.newyorkfed.org/research/staff_reports/sr514.pdf.
- Haughwout, A., Sutherland, S., Tracy, J., 2013. Negative equity and housing investment. Staff report No. 636. Federal Reserve Bank of New York (September). http://www.newyorkfed.org/research/staff_reports/2013.html.
- Haurin, D., Ma, C., Moulton, S., Schmeiser, M.D.S., Jason, S., Wei, S., 2014. Spatial variation in reverse mortgages usage: house price dynamics and consumer selection. Working paper. Ohio State University.
- Herbert, C.E., McCue, D.T., Sanchez-Moyano, R., 2014. Is homeownership still an effective means of building wealth for low-income and minority households? Was it ever? In: Belskey, E.S., Herbert, C.E., Molinsky, J.H. (Eds.), *Homeownership Built to Last*. Brookings Institution Press, Washington, DC.
- Hilber, C.A., Liu, Y., 2008. Explaining the black-white homeownership gap: the role of own wealth, parental externalities and locational preferences. *J. Hous. Econ.* 17, 152–174.
- Hilber, C.A.L., Mayer, C., 2009. Why do households without children support local public schools? Linking house price capitalization to school spending. *J. Urban Econ.* 65, 74–90.
- Hilber, C.A.L., Turner, T.M., 2014. The mortgage interest deduction and its impact on homeownership decisions. *Rev. Econ. Stat.* 96 (4), 618–637.
- Himmelberg, C., Mayer, C., Sinai, T., 2005. Assessing high house prices: bubbles, fundamentals and misperceptions. *J. Econ. Perspect.* 19, 67–92.

- HUD, 2013a. Annual Report to Congress Regarding the Financial Status of the FHA Mutual Mortgage Insurance Fund Fiscal Year 2010. U.S. Department of Housing and Urban Development.
- HUD, 2013b. Measuring the costs and savings of aging in place. Matters, Evidence.
- Hunter, W.C., Walker, M.B., 1996. The cultural affinity hypothesis and mortgage lending decisions. *J. Real Estate Financ. Econ.* 13, 57–70.
- Hurst, E., Lusardi, A., 2004. Liquidity constraints, household wealth, and entrepreneurship. *J. Polit. Econ.* 112, 319–347.
- Immergluck, D., Smith, G., 2006a. The external costs of foreclosure: the impact of single-family mortgage foreclosure on property values. *Housing Policy Debate* 17 (1), 57–79.
- Immergluck, D., Smith, G., 2006b. The impact of single-family mortgage foreclosures on neighborhood crime. *Hous. Stud.* 21, 851–866.
- Integrated Financial Engineering, 2012. Actuarial review of the Federal Housing Administration Mutual Mortgage Insurance Fund HECM loans for Fiscal Year 2012.
- Jones, L.D., 1993. Deficiency judgments and the exercise of the default option in home mortgage loans. *J. Law Econ.* 36, 115–138.
- Keenan, T.A., 2010. Home and community preferences of the 45+ population. Working paper. AARP (November). <http://assets.aarp.org/rgcenter/general/home-community-services-10.pdf>.
- Kiff, J., 2009. Canadian residential mortgage markets: boring but effective? Working paper 09/130. IMF (June).
- Ladd, H.F., 1998. Evidence on discrimination in mortgage lending. *J. Econ. Perspect.* 12, 41–62.
- Lea, M., 2010. International comparison of mortgage product offerings. Special report. Research Institute for Housing America (July).
- Lee, D., Mayer, C., Tracy, J., 2013. A new look at second liens. In: Glaeser, E.L., Sina, T. (Eds.), *Housing and the Financial Crisis*. The University of Chicago Press, Chicago, pp. 205–234.
- Lovenheim, M.F., 2011. The effect of liquid housing wealth on college enrollment. *J. Labor Econ.* 29, 741–771.
- Lubell, J., 2014. Filling the void between homeownership and rental housing: a case for expanding the use of shared equity homeownership. In: Belsky, E.S., Herbert, C.E., Molinsky, J.H. (Eds.), *Homeownership Built to Last*. Brookings Institution Press, Washington, DC.
- Lutz, B., Molloy, R., Shan, H., 2011. The housing crisis and state and local government tax revenue: five channels. *Reg. Sci. Urban Econ.* 41, 306–319.
- Mayer, C., Pence, K., 2009. Subprime mortgages: what, where, and to whom? In: Glaeser, E.L., Quigley, J. M. (Eds.), *Housing Markets and the Economy: Risk, Regulation, and Policy: Essays in honor of Karl E. Case*. Lincoln Institute of Land Policy, Cambridge, pp. 149–196.
- Mayer, C., Pence, K., Sherlund, S.M., 2009. The rise in mortgage defaults. *J. Econ. Perspect.* 23, 27–50.
- Melzer, B.T., 2012. Mortgage debt overhang: reduced investment by homeowners with negative equity. Working paper. Northwestern University, Kellogg School of Management (August). <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.412.7932&rep=rep1&type=pdf>.
- Mian, A., Sufi, A., 2009. The consequences of mortgage credit expansion: evidence from the U.S. Mortgage Default Crisis. *Q. J. Econ.* 124, 1449–1496.
- Mian, A., Sufi, A., 2011. House prices, home equity-based borrowing, and the U.S. Household Leverage Crisis. *Am. Econ. Rev.* 101, 2132–2156.
- Molloy, R., Shan, H., 2013. The post-foreclosure experience of U.S. Households. *Real Estate Econ.* 41, 225–254.
- Morgan, D., Iverson, B., Botsch, M., 2012. Subprime foreclosures and the 2005 bankruptcy reform. *Econ. Pol. Rev.* 18, 47–57.
- Moulton, S., Haurin, D.R., Shi, W., 2014. An analysis of default risk in the home equity conversion mortgage (HECM) program. Working paper. Ohio State University. <http://ssrn.com/abstract=2468247>.
- Munnell, A.H., Tootell, G.M.B., Browne, L.E., McEneaney, J., 1996. Mortgage lending in Boston: interpreting HMDA data. *Am. Econ. Rev.* 86, 25–53.
- National Association of Realtors, 2013. Housing pulse survey, Washington, DC. <http://www.realtor.org/reports/housing-pulse-surveys>.

- Nenov, P., 2010. Labor market and regional reallocation effects of housing busts. Working paper. MIT.
- OCC mortgage metrics report: fourth quarter 2013. Comptroller of the Currency, US Department of the Treasury, Washington, DC (December).
- Office of Management and Budget, 2014. Analytical perspectives. Government Printing Office.
- O'Flaherty, B., 1993. Abandoned buildings: a stochastic analysis. *J. Urban Econ.* 34, 43–74.
- Palmer, C., 2014. Why did so many subprime borrowers default during the crisis: loose credit or plummeting prices. Working paper. MIT (January).
- Pinto, E., 2012. How the FHA hurts working class families and communities. AEI.
- Poterba, J., Sinai, T., 2008. Tax expenditures for owner-occupied housing: deductions for property taxes and mortgage interest and the exclusion of imputed rental income. *Am. Econ. Rev.* 98, 84–89.
- Poterba, J., Venti, S., Wise, D., 2011. The composition and drawdown of wealth in retirement. *J. Econ. Perspect.* 25, 95–118.
- Quigley, J.M., 1987. Interest rate variations, mortgage prepayments and household mobility. *Rev. Econ. Stat.* 69, 636–643.
- Rao, J., 2012. The Other Foreclosure Crisis: Property Tax Lien Sales. National Consumer Law Center, Boston, MA (July 14). <https://www.nclc.org/issues/the-other-foreclosure-crisis.html>.
- Ross, S., Yinger, J., 2002. *The Color of Credit: Mortgage Lending Discrimination, Research Methodology, and Fair Lending Enforcement*. MIT Press, Cambridge, MA.
- Ross, S.L., Turner, M.A., Godfrey, E., Smith, R.R., 2008. Mortgage lending in Chicago and Los Angeles: a paired testing study of the pre-application process. *J. Urban Econ.* 63, 902–919.
- Saiz, A., 2008. On local housing supply elasticity. Working paper. University of Pennsylvania. http://papers.ssrn.com/so13/papers.cfm?abstract_id=1193422.
- Scafidi, B.P., Schill, M.H., Wachter, S.M., Culhane, D.P., 1998. An economic analysis of housing abandonment. *J. Hous. Econ.* 7, 287–303.
- Schill, M.H., Wachter, S.M., 1993. A tale of two cities: racial and ethnic geographic disparities in home mortgage lending in Boston and Philadelphia. *J. Hous. Res.* 4 (2), 245–275.
- Schuetz, J., Been, V., Ellen, I.G., 2008. Neighboring effects of concentrated mortgage foreclosures. *J. Hous. Res.* 17, 306–319.
- Schulhofer-Wohl, S., 2011. Negative equity does not reduce homeowners' mobility. Working paper No. 16701. NBER (January). <http://www.nber.org/papers/w16701>.
- Shan, H., 2011. Reversing the trend: the recent expansion of the reverse mortgage. *Real Estate Econ.* 39, 743–768.
- Shiller, R., 2008. *The Subprime Solution: How Today's Global Financial Crisis Happened, and What to Do About It*. Princeton University Press, Princeton, NJ.
- Shiller, R., 2014. Why is housing finance still stuck in such a primitive stage? *Am. Econ. Rev.* 104, 73–76.
- Stein, J., 1995. Prices and trading volume in the housing market: a model with down-payment effects. *Q. J. Econ.* 110, 379–406.
- Tootell, G.M.B., 1996. Redlining in Boston: do mortgage lenders discriminate against neighborhoods? *Q. J. Econ.* 111, 1049–1079.
- Tractlet, V., 2010. An overview of the Canadian housing finance system. *Housing Finance International*, 6–13.
- Turner, M.A., Santos, R., Levy, D.K., Wissoker, D., Aranda, C., Pitingolo, R., 2013. *Housing Discrimination Against Racial and Ethnic Minorities 2012*. Research report. U.S. Department of Housing and Urban Development, Office of Policy Development (June).
- Valletta, R.G., 2013. House lock and structural unemployment. *Labour Econ.* 25, 86–97.
- White, M.J., 1986. Property taxes and urban housing abandonment. *J. Urban Econ.* 20, 312–330.
- Wilson, E., Callis, R.R., 2013. Who could afford to buy a home in 2009? Working paper. <http://www.census.gov/prod/2013pubs/h121-13-02.pdf>.
- Woodward, S.E., Hall, R.E., 2012. Diagnosing consumer confusion and sub-optimal shopping effort: theory and mortgage-market evidence. *Am. Econ. Rev.* 102, 3249–3276.
- Zhu, J., 2012. An empirical analysis of the home affordable refinance program's impact on default rates. Working paper. Freddie Mac (May).

This page intentionally left blank

CHAPTER 16

Change and Persistence in the Economic Status of Neighborhoods and Cities

Stuart S. Rosenthal*, Stephen L. Ross[†]

*Maxwell Advisory Board Professor of Economics, Department of Economics, Syracuse University, Syracuse, NY, USA

[†]Department of Economics, University of Connecticut, Storrs, CT, USA

Contents

16.1. Introduction	1048
16.2. Neighborhood Economic Status	1054
16.2.1 Four stylized facts about neighborhood economic status	1054
16.2.1.1 <i>Rich suburbs and poor cities</i>	1054
16.2.1.2 <i>Change in neighborhood economic status is common</i>	1055
16.2.1.3 <i>Mean reversion in neighborhood economic status is also common</i>	1056
16.2.1.4 <i>Income segregation has increased in US cities</i>	1058
16.2.2 Conceptual models of spatial variation in neighborhood economic status	1058
16.2.2.1 <i>Static models</i>	1058
16.2.2.2 <i>Dynamic models</i>	1066
16.2.3 Evidence of drivers of neighborhood economic status	1074
16.2.4 The pace and timing of change in neighborhood economic status	1077
16.2.4.1 <i>Home maintenance and depreciation of housing capital</i>	1078
16.2.4.2 <i>The decision to redevelop</i>	1081
16.2.4.3 <i>Filtering and gentrification</i>	1084
16.2.5 Does gentrification hurt local residents?	1086
16.2.6 US context versus Europe	1088
16.3. City Dynamics	1088
16.3.1 Overview	1088
16.3.2 Three previous stylized facts and one new observation	1089
16.3.2.1 <i>Stable city size distributions that approximate the rank-size rule</i>	1089
16.3.2.2 <i>Slow movement of individual cities up and down the city size distribution</i>	1090
16.3.2.3 <i>Fast changes in city industrial mix relative to total employment</i>	1090
16.3.2.4 <i>Very fast movement of cities up and down in economic status based on income</i>	1091
16.3.3 Drivers of persistence in metropolitan economic status	1094
16.3.3.1 <i>Natural advantages</i>	1094
16.3.3.2 <i>Agglomeration economies</i>	1096
16.3.3.3 <i>Culture</i>	1097
16.3.4 Drivers of change in metropolitan economic status	1098
16.3.4.1 <i>Productivity and labor demand</i>	1098
16.3.4.2 <i>Amenities and labor supply</i>	1102
16.3.4.3 <i>Inelastic housing supply</i>	1104

16.4. Conclusions and Future Research	1106
Appendix Supplemental Figures	1108
Acknowledgments	1114
References	1114

Abstract

This chapter reviews recent literature that considers and explains the tendency for neighborhood and city-level economic status to rise and fall. A central message is that although many locations exhibit extreme persistence in economic status, change in economic status as measured by various indicators of *per capita* income is common. At the neighborhood level, we begin with a set of stylized facts and then follow with discussion of static and dynamic drivers of neighborhood economic status. This is mirrored at the metropolitan level. Durable but slowly decaying housing, transportation infrastructure, and self-reinforcing spillovers all influence local income dynamics, as do enduring natural advantages, amenities, and government policy. Three recurring themes run throughout the paper: (i) Long sweeps of time are typically necessary to appreciate that change in economic status is common, (ii) history matters, and (iii) a combination of static and dynamic forces ensures that income dynamics can and do differ dramatically across locations but in ways that can be understood.

Keywords

Neighborhood income dynamics, City income dynamics, Durable housing, Transportation infrastructure, Spillovers, Persistence, Path dependence, Cycles

JEL Classification Codes

R0, R1, R2, R3, R4

16.1. INTRODUCTION

At both the neighborhood and metropolitan levels, sharp differences in economic status are common. Within US metropolitan areas, for example, between 2005 and 2009, the average difference in median income between the 75th percentile census tract and the 25th percentile census tract was 54.8%.¹ Across US metropolitan areas, the difference in MSA (metropolitan statistical area) median income between the 75th percentile MSA and the 25th percentile MSA was 24.5%.² The extensive level of income inequality across locations implicit in these measures is well known and in many instances seemingly entrenched. It is hard to imagine Back Bay, Boston, for example, with its famous brownstones losing its appeal to high-income residents, or San Francisco with its natural amenities.³

¹ The authors' calculation based on the 366 largest CBSA (core-based statistical areas) in the United States using pooled 1% samples from the 2005 to 2009 American Community Survey (ACS).

² MSA refers to metropolitan statistical area. The estimate above is based on the authors' calculation using the 1% sample of the 2010 ACS.

³ See the Back Bay Neighborhood Association website for a history of Back Bay, which, until the landfill project begun in 1857, was a tidal marsh: http://www.nabbonline.com/about_us/back_bay_history.

Nevertheless, *change* in a location's economic status is common. Rosenthal (2008a), for example, showed that over half of all census tracts in the core areas of 35 US cities were of markedly different economic status in 2000 relative to their status in 1950. The rise, fall, and rise again of Harlem over the last 100 years are emblematic of this change. At the metropolitan level, Glaeser and Gyourko (2005) emphasized that 11 of the 15 largest US cities in 1950 lost population by 2000; most dramatic among these, Detroit lost over half of its population!

A central goal of this chapter is to review recent literature that provides new insights into the extent and manner in which a location's economic status tends to change over time, at both the neighborhood and MSA levels. When focusing on neighborhood change, much of our discussion will be set in a US context as this has been the tendency in the literature. Nevertheless, the lessons learned from that literature apply to cities beyond the United States. When focusing on metropolitan-level change, the literature is more international as is our review. In most instances, we will refer to various measures of *per capita* income to characterize local economic status, but in some instances, we will refer to changes in population and employment.⁴

Throughout the chapter, a recurring theme is that long sweeps of time are necessary to appreciate that change in a location's economic status is common. A related theme is that history matters and in some instances contributes to path dependency. A third message is that multiple forces affect neighborhood and MSA-level economic status so that income dynamics differ across locations. As a result, in some locations, economic status exhibits extreme persistence, while in other locations, economic status is much more prone to random and systematic forces that prompt cycles of rise and fall.

Three mechanisms figure prominently in our discussion and often reinforce each other. These include the durability and slow decay of housing, the durability of transportation infrastructure, and self-reinforcing spillovers that have the potential to root ethnic/racial communities to a given neighborhood and seemingly footloose industries to a given location. In many instances, these mechanisms contribute to slow, relentless change that extends over decades. In other cases, economic forces build up until a tipping point is reached after which the pace of change can be dramatic.

At the neighborhood level, it is also important to recognize that income displays sharp, nonrandom spatial patterns. This is graphically apparent in Figure 16.1, which presents plots of neighborhood relative income as a function of distance from the city center for four metro areas in the United States over the 2005–2009 period. The plots include New York, (Panel 1), Los Angeles (Panel 2), Chicago (Panel 3), and Philadelphia

⁴ Our emphasis on *per capita* income is different from previous reviews of MSA-level productivity and growth, which tend to focus on population and employment (e.g., Rosenthal and Strange, 2004; Behrens and Robert-Nicoud, this volume; Combes and Gobillon, this volume; Carlinio and Kerr, this volume).

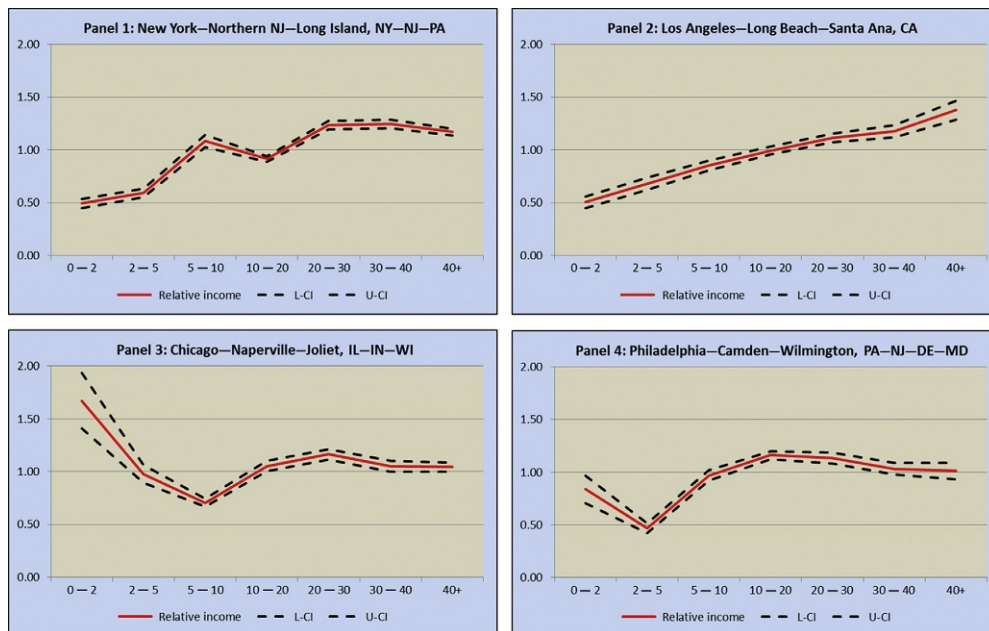


Figure 16.1 Census tract median income relative to its MSA by distance (in miles) to the MSA center (2005–2009 ACS pooled sample).

(Panel 4). For each metro area, neighborhood relative income is measured based on census tract average income divided by the tract's MSA average income. In Los Angeles, relative income rises in a strikingly monotonic fashion with distance from the center. Relative income also trends upward with distance from the downtown in New York, Chicago, and Philadelphia, but in a much less monotonic fashion, a point that we will return to. Overall, these four cities are indicative of a widely documented stylized fact: US central cities tend to be poor relative to the suburbs.

The broad spatial features in Figure 16.1 have been the focus of intense study since the late 1960s. Much less well appreciated is that a location's economic status is not static. This is evident in Figure 16.2, which displays summary measures drawn from table 1 of Rosenthal (2008a). Four vertical bars are displayed in the figure, each of which is broken into four segments with 1950 census tract economic status increasing to the right and year 2000 economic status increasing vertically. For these purposes, status is measured by the ratio of a census tract's average income relative to its city.

Figure 16.2 indicates that regardless of whether a census tract was of low- or high-income status in 1950 (the first and fourth bars, respectively), over half of all census tracts in 1950 were of markedly different economic status relative to their respective cities in 2000. Moreover, Rosenthal (2008a) showed that this pattern holds regardless of whether the census tract is situated in a large or a small city. The decline of several of the largest US

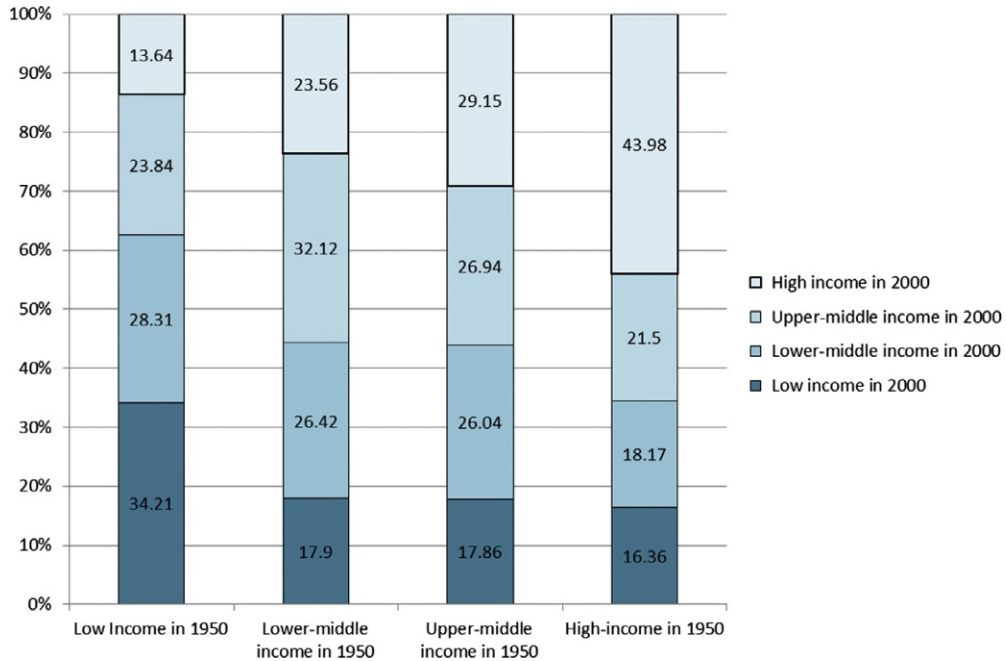


Figure 16.2 Transition rates of census tract relative income between 1950 and 2000. Based on census tracts from 35 cities followed on a consistent geographic basis from 1950 to 2000. Tract relative income is measured as census tract average income relative to the average income among tracts in the MSA in question and that are contained in the balanced panel. Source: [Rosenthal \(2008a\)](#), Panel (a) of table 1, p. 2.

cities as noted by [Glaeser and Gyourko \(2005\)](#) above is equally striking. These changes, however, require time to observe. Rosenthal reported that the average change in census tract income relative to its MSA is roughly 12% per decade (in absolute value). For the 10 largest cities in the United States in 1977, [Duranton \(2007\)](#) reported that the average change in population size rank over the following 20 years was just 1.2 from among 272 metropolitan areas (also in absolute value). [Findeisen and Südekum \(2008\)](#) documented similar rates of change for German metropolitan areas. At this pace of change, if one were to adopt a time horizon of, say, just one decade, it would be easy to overlook the enormously dynamic nature of neighborhoods and cities.

Nevertheless, that is often what policy makers and a vast academic literature on neighborhood and metropolitan productivity and growth have tended to do, consistent with our natural tendency to focus on the near term. Forty-five percent of all renters in the United States remain in their homes for less than 2 years, while a similar fraction of owner-occupiers move in under 9 years. In 2010, 14.1% of these moves were to a

different state, 16.8% were to a different PUMA within state, and 66.3% were within the same PUMA.⁵ These sorts of mobility rates and relocation patterns help to explain the tendency to focus on the short term, at least when considering policies that affect the economic vitality of local communities.⁶

Two examples will help to frame our discussion as we review literature relevant to these ideas, Harlem and Detroit. With the extension of the New York subway system out to Harlem in 1904, local real estate developers expected a boom in housing demand and built homes in anticipation of incoming families.⁷ Much of that demand did not materialize because of concurrent improved access to downtown Manhattan from other locations. Local folklore has it that Phillip Payton, a real estate agent and entrepreneur, approached several Harlem landlords and convinced them to encourage black families to move to Harlem to occupy the new housing stock. By the 1920s, Harlem had transformed into a thriving black community, rich in culture, jobs, and hope. That promise, however, was dealt a blow with the Great Depression of the 1930s, discrimination, and social upheaval. Harlem's population shrank in the late 1960s and 1970s as crime, riots, abandoned property, and blight dominated the headlines instead of iconic sites like the Cotton Club and the Apollo Theater. By 1990, Harlem had become an impoverished black community living in aging and often decrepit housing.⁸ This too changed.

Where abandoned and rundown buildings were found in 1990, newly built homes now sell for over 1 million dollars and count Bill Clinton's business office as a neighbor. From 2000 to 2008, Central Harlem's population rose from 109,000 to 126,000. These changes have also coincided with the most profound change in the composition of Harlem's population in 100 years. Blacks in greater Harlem are no longer a majority but instead account for roughly 40% of the population, down from a high of roughly 64% in 1970. Central Harlem's black population share is said to have peaked at 98% in 1950 but stood at just 62% in 2008.⁹ These changes reflect both out-movement of blacks—Central Harlem's black population in 2008 had fallen to 77,000, a level not seen since the 1920s—and in-movement of large numbers of Hispanic and non-Hispanic whites.¹⁰ We will argue that a variety of studies on neighborhood dynamics suggest that Harlem's rise and fall and its transition from white to black and back again are reflective of

⁵ The authors' calculations based on the 2010 American Community Survey. For renters, 14.2% of moves are to a different state, 16.3% are to a different public use microdata area (PUMA) within state, and 66.9% are within PUMA. For owner-occupiers, the corresponding numbers are 13.8%, 18.5%, and 64.3%.

⁶ Electoral politics also reinforce the tendency to focus on the near term as incumbents and challengers compete for votes from an electorate that discounts events in the future.

⁷ See "History of Harlem Heritage Tours & Cultural Center." *Harlem Heritage Tours Cultural Center RSS*. See also Vincent (2005). "Harlem." *Continuum Encyclopedia of Popular Music of the World: Locations*.

⁸ "History of Harlem Heritage Tours & Cultural Center." *Harlem Heritage Tours Cultural Center RSS*.

⁹ Payne (2010) "Is Harlem No Longer Black?" *The Root*, 8 Jan.

¹⁰ Roberts (2010) "No Longer Majority Black, Harlem Is in Transition." *The New York Times*, 5 January.

many urban neighborhoods in the United States and elsewhere in the world. We believe these dynamics reflect the combined forces of access to transportation infrastructure, durable but slowly depreciating housing, and the self-reinforcing cultural effects of race and ethnicity. Importantly, a conclusion we draw from the literature is that the sort of changes experienced by Harlem are inevitable for many—but not all—urban communities.

In 1950, Detroit was the fifth largest city in the United States with a population of roughly 1.85 million people, 45% of whom were white, while the rest were almost exclusively black. By 2010, Detroit's population had fallen to just over 700,000 people and whites accounted for only 7.9% of the city's population. The city officially succumbed in July 2013 when it filed for Chapter 9 bankruptcy, the largest municipal bankruptcy (by debt) in the history of the United States. Detroit's dramatic decline is emblematic of many declining US rust belt cities that include Buffalo, Pittsburgh, Cleveland, and others. Shrinking cities are also found outside of the US rust belt as with Liverpool, England, whose population in 2010 was down nearly 50% from its peak in the 1930s, and Torino, Italy, which lost roughly 25% of its population between 1970 and 1990.¹¹ Other cities have disappeared altogether, as with Memphis, Egypt, which was once a leading city in the world.¹² These cities remind us that metropolitan areas can shrink as well as grow.

As with our assessment of neighborhood dynamics, durable housing, transportation infrastructure, and fiscal policies played an important role in both the growth and demise of Detroit. Ease of transportation provided access to raw materials as well as distant markets and suppliers, enhancing trade, productivity, and growth. Steel is an essential input for the auto industry, and the raw materials necessary for steel production are found in vast quantities in areas close to the Great Lakes. This contributed to the emergence of Gary, Indiana, and Pittsburgh, Pennsylvania, as steel-producing towns, which in turn contributed to the rise of nearby Detroit as "Motor City."

Racial strife, fiscal mismanagement, and especially the declining comparative advantage of Detroit as a world automobile center have since contributed to six decades of steady decline in Detroit's employment base. The same durable buildings that helped to support Detroit's growing population during its rise deepened and accelerated its fall

¹¹ See, for example, Nurse (2008) in the University of Liverpool, *University News*, "The Liverpool View: Detroit's Lessons for Liverpool" at <http://news.liv.ac.uk/2013/08/02/the-liverpool-view-detroits-lessons-for-liverpool/>. Power et al. (2008) also document and discuss population decline in seven European cities between 1970 and 1990, including Leipzig (−15%), Bremen (−5%), Sheffield (−8%), Bilbao (−13%), Torino (−25%), and Saint Etienne (−20%) (see Chart 6, p. 11 of their report).

¹² Memphis was founded roughly 5000 years ago adjacent to the southern end of the Nile Delta (Prasad (1977)). In its heyday from around 3000 BC to around 2250 BC, Memphis is thought to have been the largest city in the world, with a population that may have been as high as 50,000 people (Chandler, 1987). Nevertheless, where Memphis stood, today, one finds scattered ruins and small villages (Waters, 1895).

as it declined. That is because reduced demand in conjunction with inelastic housing supply causes home prices to plummet, which reinforce the shift toward a low-skill, lower-income population (e.g., Glaeser and Gyourko, 2005). Declining skill simultaneously lowers *per capita* income and reinforces the tendency for employers to flee the city. These shifts erode the local tax base and aggravate the city's growing fiscal crisis. In response, many declining US rust belt cities including Buffalo, Cleveland, and especially Detroit have adopted policies of bulldozing underutilized and abandoned housing stock in an attempt to *strengthen* their local economies. It is in that spirit that Glaeser (2007) concluded that Buffalo should "shrink to greatness."

The history of rust belt towns in the United States suggests that cities can lose their comparative advantages as documented by Yoon (2013), after which durable housing stocks contribute to a sharp downward spiral. The idea that cities can shrink and even disappear, however, has been almost overlooked in the literature on urban dynamics and growth. It also contrasts with recent work by Davis and Weinstein (2002, 2008), Brakman et al. (2004), and Glocker and Sturm (2013), all of whom suggest that enduring locational natural advantages help to ensure that a city's economic status remains secure even following the devastation of war. The view that cities can not only rise but also fall challenges recent literature that suggests that once built, a city will endogenously grow and endure even after losing its initial comparative advantage, as with the obsolescence of canoe portage paths (e.g., Bleakley and Lin, 2012).

To explore these ideas, our chapter is organized in a manner that progresses from neighborhood-level to MSA-level geography. We begin by describing four stylized facts that pertain to the location of high- and low-income communities within individual cities. This is followed by a series of static and dynamic arguments that explain patterns documented in recent studies in addition to Harlem's experience highlighted earlier. An important conclusion is that while random shocks contribute to a community's economic status, in most instances, neighborhood economic status follows from a mix of systematic static and dynamic economic forces. From there, we review papers that consider change at the metropolitan level. Here too, we outline four stylized facts and focus on whether a city's economic status relative to other metropolitan areas is fixed or whether it changes over time in a systematic fashion, and why. We conclude by highlighting areas that we believe are especially in need of further research.

16.2. NEIGHBORHOOD ECONOMIC STATUS

16.2.1 Four stylized facts about neighborhood economic status

16.2.1.1 *Rich suburbs and poor cities*

A prominent feature of modern cities in the United States has been the tendency for suburban communities to have high levels of income relative to the central cities. Glaeser et al. (2008), Brueckner and Rosenthal (2009), and others had confirmed this. We do

so again here using pooled census tract data from the ACS from 2005 to 2009. Los Angeles provides a graphic example as shown in Panel 2 in [Figure 16.1](#), with household income relative to the MSA rising monotonically with distance (in miles) from the MSA center.¹³ The pattern for New York, which is shown in Panel 1, is similarly upward sloping although with more of a sawtooth shape to the plot. Given the striking patterns for Los Angeles and New York, and decades of concern about the concentration of poverty in the inner cities (e.g., [Rosenthal, 2008b](#)), it is easy to be lulled into thinking that suburban household incomes must nearly always be high relative to their MSA centers, both now and in the future. Nevertheless, one can point to instances where that is not the case, the most dramatic of which in the United States is Chicago. In Panel 3 of [Figure 16.1](#), notice that for Chicago, central city income is high relative to the MSA, falls to a trough at roughly 5–10 miles from the city center, and then rises again. Philadelphia (shown in Panel 4 of [Figure 16.1](#)) also displays a pronounced v-shaped contour although its central city is still of relatively low-income status. Similar plots are provided in Appendix for the 48 largest MSAs in the United States. While it is clear that high-income suburban areas are the norm, it is also clear that there are other cities in addition to Chicago and Philadelphia where that is not the case as with Washington, DC, to name just one (see Panel 7 of [Figure A1](#)). Nevertheless, although there are important exceptions, our first stylized fact is that in most metropolitan areas of the United States, the suburbs are of higher-income status, while the central cities are relatively poor.

16.2.1.2 Change in neighborhood economic status is common

It is tempting to assume that a neighborhood's economic status as high or low income is largely fixed in place but that would not be correct. [Rosenthal \(2008a\)](#) calculated transition rates of neighborhoods between different levels of economic status using a balanced panel of census tracts for core areas of 35 cities that are followed on a consistent geographic basis from 1950 to 2000. Each census tract is treated as a separate neighborhood. Status is measured based on average household income in a neighborhood relative to average household income of all census tracts in the panel for the city and year in which the neighborhood is observed. Neighborhood geography is coded to year 2000 census tract boundaries for all years. Neighborhoods are further classified into four groups based on whether neighborhood relative income levels are in the first through fourth quartiles of relative income, referred to by [Rosenthal \(2008a\)](#) as low-income, lower middle-income, upper middle-income, and high-income, respectively.

A striking pattern emerges and is reproduced in [Figure 16.2](#). For the 35 cities in the sample, only 34.21% of all low-income neighborhoods in 1950 were still of low-income

¹³ In each of the panels of [Figure 16.1](#), distance from the most densely populated census tract is plotted along the horizontal axis, while the average ratio of census tract median income to the MSA median is on the vertical axis. Data for these plots are from the combined 2005–2009 American Community Survey.

status in 2000. For lower-middle-income tracts, upper-middle-income tracts, and high-income tracts, the corresponding values are 26.42%, 26.94%, and 43.98%, respectively. Although there is a disproportionate tendency for low- and high-income tracts to remain as such, the overwhelming pattern is that most neighborhoods change economic status between 1950 and 2000, either up or down the economic ladder.

These patterns point to our second stylized fact: Change in neighborhood economic status is common among urban neighborhoods in the United States. Yet most of our models and much of the urban policy seem to implicitly treat neighborhood economic status as stationary.

16.2.1.3 Mean reversion in neighborhood economic status is also common

With so much change in neighborhood economic status, a natural question is whether neighborhoods rise and fall in a systematic, cyclical pattern or whether individual neighborhood economic status follows a random walk. Table 16.1 reproduces estimates from Rosenthal (2008a, table 3, p. 5) that shed light on this question. The table reports regressions that characterize the degree to which change in neighborhood economic status is serially correlated. As above, neighborhood economic status (y_{it}) for neighborhood i in

Table 16.1 Serial correlation in growth in neighborhood relative income (t-ratios based on robust standard errors in parentheses)

	Philadelphia County ward panel ^a	Philadelphia County ward panel ^a	Philadelphia County census tract Balanced panel ^b	35 MSA Census tract balanced panel ^b
	$\log(y_{2000}/y_{1950})$	$\log(y_{1980}/y_{1950})$	$\log(y_t/y_{t-1})$	$\log(y_t/y_{t-1})$
$\log(y_{1950}/y_{1900})$	-0.9465 (-6.17)	—	—	—
$\log(y_{1950}/y_{1920})$	—	-0.4538 (-3.20)	—	—
$\log(y_{t-1}/y_{t-2})$	—	—	-0.0126 (-0.24)	-0.0564 (-11.27)
Constant	-0.0617 (-1.10)	-0.0423 (-0.82)	-0.0863 (-10.66)	—
Period length (years)	50	30	10	10
Time span	1900–2000	1920–2000	1950–2000	1950–2000
County fixed effects	—	—	—	125
Observations	39	39	1304	37,676
R-square	0.4329	0.1749	0.0002	0.0323

^aAll data were coded to year 1900 ward boundaries. See Rosenthal (2008a) for details.

^bAll data were coded to year 2000 census tract boundaries. See Rosenthal (2008a) for details.

Source: Rosenthal (2008a, based on table 3, p. 5).

period t is measured as the ratio of average income in the neighborhood (census tract) relative to average income in its MSA.¹⁴ For 1900 and 1920, income in each voting ward is based on the OCCSCORE measure available in IPUMs, which provides an estimate of income that an individual would have earned in 1950 given their actual occupation in 1900 or 1920 (see www.ipums.org for details). Income after 1920 is based on actual reported individual income, and for all years, individual income is aggregated to the geographic unit used in the regression (e.g., 1900 voting ward boundaries or census tracts).

The first two regressions focus only on Philadelphia County with data coded to year 1900 voting ward geography. There were 39 such wards in Philadelphia in 1900. The period length for the first regression is set to 50 years with $\log(\gamma_{2000}/\gamma_{1950})$ regressed on $\log(\gamma_{1950}/\gamma_{1900})$. In the second regression, the period length is set to 30 years with $\log(\gamma_{1980}/\gamma_{1950})$ regressed on $\log(\gamma_{1950}/\gamma_{1920})$. The third regression in the table also focuses on Philadelphia County, but in this instance, data are coded to year 2000 census tracts. In addition, census tracts are followed from 1950 to 2000 on a decade-by-decade basis with $\log(\gamma_t/\gamma_{t-1})$ regressed on $\log(\gamma_{t-1}/\gamma_{t-2})$. The fourth and last regression in the table is similarly specified but includes census tracts for the core areas of 35 MSAs for which tracts were defined in 1950. This latter regression also includes county fixed effects.

Moving from left to right in Table 16.1 across the Philadelphia County regressions (in columns 1–3), notice that the coefficients on the lagged-dependent variables are -0.95 , -0.45 , and -0.013 , respectively, with t -ratios of 6.2, 3.2, and 0.24. This indicates that for the typical neighborhood in Philadelphia, a neighborhood's relative economic status in 2000 is 95% back to where it began 100 years earlier in 1900. As the period length narrows in the second and third regressions, neighborhood cycles are less complete and the coefficients on the lagged-dependent variables are reduced, falling to 1.3% for the 10-year period length model. This latter estimate is close to the corresponding measure obtained for the 35-city balanced panel in the last regression. In that model, the coefficient on the lagged-dependent variable is -5.6% (with a t -ratio of 11.27).

An implication of these results is that neighborhood income is stationary, which can be tested using panel unit root tests. Consider the following equation:

$$\log(\gamma_{it}) = \theta_{i,0} + \theta_{i,1} \log(\gamma_{i,t-1}) + e_{it}, \quad (16.1)$$

where $\log(\gamma_{it})$ is expressed as a function of a constant and its one period lag and i and t denote the census tract and time period as before. If $\theta_{i,1} < 1$, $\log(\gamma_{it})$ is stationary with finite variance and a stable long-run mean. This would imply that neighborhood

¹⁴ To be precise, let γ_{it} be the relative income of neighborhood i ($i = 1, \dots, J$) in period t . In addition, γ_{it} is defined to be $Y_{it}/\bar{Y}_t$, where Y_{it} is the average level of income in tract i in period t , while $\bar{Y}_t$ is the city-wide average level of income in period t . By construction, the expected value of γ over all neighborhoods in period t equals 1.

economic status displays a tendency for mean reversion. If instead θ_1 equals 1, then shocks to neighborhood economic status never damp out and neighborhood economic status would follow a random walk.

Using the same data as in column 4 of Table 16.1, Rosenthal (2008a) estimated panel unit root tests that clearly reject the null of a unit root in favor of the view that neighborhoods vary around a stable long-run mean.¹⁵ These results lend support to the idea that Harlem's 100-year cycle of rise, fall, and renewal described in Section 16.1 is not unique and that urban neighborhoods tend to exhibit long-running cycles in economic status and mean reversion.

16.2.1.4 Income segregation has increased in US cities

A series of studies provide evidence that income segregation has increased within US metropolitan areas in recent decades. Massey and Fischer (2003) reported evidence of regional income convergence between 1950 and 2000; they also found that inequality in income and poverty across census tracts increased between 1970 and 1990 while declining only slightly between 1990 and 2000. Taylor and Fry (2012) reported an increase in census tract-level income segregation between 1980 and 2010, which was especially pronounced among higher-income households. Wheeler and La Jeunesse (2007) found that income segregation across census tracts within a given US MSA increased substantially in the 1980s. Watson (2009) presented measures of within metropolitan income segregation over time based on income rank percentile as opposed to thresholds tied to real income levels. Watson also reported that within-MSA income inequality increased between 1970 and 2000 with most of the increase occurring between 1980 and 1990.

16.2.2 Conceptual models of spatial variation in neighborhood economic status

The evidence above suggests that neighborhood economic status often exhibits pronounced systematic spatial and temporal patterns but does not explain why. This section offers a series of conceptual explanations that help to clarify the drivers of the patterns above.

16.2.2.1 Static models

16.2.2.1.1 The "standard" model: Alonso (1964), Mills (1967), and Muth (1969)

Early work by Alonso (1964), Mills (1967), and Muth (1969), hereafter AMM, provided the foundations for the first seemingly compelling explanation for why modern US central cities tend to have low income relative to the suburbs. In its simplest form, all

¹⁵ The two tests were those of Levin et al. (2002) referred to as LLC and a Fisher-type method developed by Maddala and Wu (1999).

employment in a metropolitan area is concentrated in the central city (i.e., the metropolitan area is “monocentric”) so that commuting costs increase with distance from the city center (u) at rate t . Households have income (y), and value housing (h) and non-housing consumption (x) for which the per unit prices are $P_h(u)$ and 1, respectively. With identical households, a spatial equilibrium would require that house prices fall with distance to the downtown to compensate for differences in commuting costs:

$$\frac{\partial P_h(u)}{\partial u} = -\frac{t(y)}{h(y)} < 0, \quad (16.2)$$

where unit commuting costs and housing demand vary with income.¹⁶ The impact of income on the slope of the house price function is then given by

$$\frac{\partial^2 P_h(u)}{\partial u \partial y} = c(y) [\varepsilon_{h,y} - \varepsilon_{t,y}], \quad (16.3)$$

where $c(y) = \frac{t(y)}{yh(y)} > 0$ and $\varepsilon_{h,y}$ and $\varepsilon_{t,y}$ are the income elasticities of demand for housing and of commuting cost, respectively.

Expression (16.3) yields the important result that if $\varepsilon_{h,y} > \varepsilon_{t,y}$, the house price function flattens with an increase in income, while the opposite is true if $\varepsilon_{h,y} < \varepsilon_{t,y}$. Drawing on this principle, through most of the 1970s, 1980s, and 1990s, a common explanation offered for why US central cities tend to be poor was that $\varepsilon_{h,y} > \varepsilon_{t,y}$. In that event, the house price functions for high- and low-income households cross as in Figure 16.3, and high-income families outbid the poor for space in the suburbs, while the poor outbid the rich for space in the central cities.

The AMM model, it seemed, offered an elegant explanation for the spatial pattern of high- and low-income neighborhoods, and this became part of the standard discussion in many classes on urban economics. An early voice of dissent, however, was raised by Wheaton (1977) who questioned whether housing demand is actually more sensitive to income than are commuting costs. Wheaton’s critique was largely overlooked until being revisited by Glaeser et al. (2008), or GKR, who provided compelling evidence that housing demand is far less sensitive to income as compared to commuting costs.

GKR emphasized that if households value commuting time in a manner that is proportional to the household head’s hourly wage, then the income elasticity of commuting costs will equal 1. Drawing on a sample of households from the American Housing Survey (AHS), GKR also estimated the income elasticity of demand for land among

¹⁶ Expression (16.2) is obtained by rearranging the household budget constraint to ensure that house prices vary in a manner that compensates households for longer commutes:

$$y = P_h(u)h(y) + x(y) + t(y)u \rightarrow P_h(u) = \frac{y - x(y)}{h(y)} - \frac{t(y)}{h(y)}u.$$

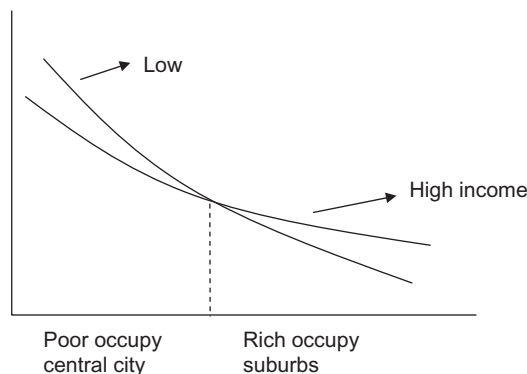


Figure 16.3 AMM and income stratification with $\epsilon_{h,y} > \epsilon_{t,y}$.

homeowners occupying single-family detached homes. The income elasticity of demand for land is roughly 8% when estimating by OLS and 25% when income is instrumented using education of the household head.¹⁷ Rosenthal (2014) recently estimated the income elasticity of demand for housing (not land) using all households other than those living in mobile homes. Based on OLS estimates, Rosenthal (2014) reports an income elasticity of 41% for owner-occupiers and 12% for renters. Together, these and numerous similar estimates in the literature confirm that the income elasticities of demand for housing and land are well below 1.

Armed with their evidence, GKR argued that the AMM model and expression (16.3) above predict the opposite pattern from what is typically observed in the United States: specifically, higher-income families should occupy the city centers, not the suburbs. A new explanation was needed for the spatial pattern of high- and lower-income communities in US metropolitan areas.

16.2.2.1.2 Public transit and access to cars

In seeking to explain the discrepancy above, Glaeser et al. (2008) emphasized that lower-income families own fewer cars per adult and are more reliant on public transit for that reason. Moreover, public transit tends to be far more effective in densely developed central city areas where demand is high enough to substantively reduce average fixed costs and also allow for more frequent service. For these reasons, public transit opportunities

¹⁷ Housing demand is based on forward-looking expectations of future income and for that reason is in principle more closely tied to permanent as opposed to current income. GKR instrumented for income using education of the household head to allow for this distinction and obtain the anticipated result that income elasticity is higher relative OLS estimates that include current income as the primary control. GKR also recognized that education could potentially enter directly into the housing demand function and for that reason may not be a valid instrument. Nevertheless, the estimates obtained are far below 1 and reinforce the arguments above.

will generally be more accessible in the central cities than in the suburbs and that should attract lower-income families to the center.

GKR documented that public transit usage is generally notably lower as one moves away from the city center. The primary exception is that in metropolitan areas with subway systems, public transit usage increases in the first few miles from the city center and then declines (see figure 3 of GKR). We reconfirm the core features of these patterns here. As in [Brueckner and Rosenthal \(2009\)](#), each census tract is said to provide good or bad access to public transit (1 or 0, respectively) depending on whether 10% or more of households in the census tract use public transit as their primary means of travel to work. Transit access is then averaged across tracts within a given distance band (e.g., 1 mile from the center or 1–2 miles from the center). For the four metro areas highlighted above (New York, Los Angeles, Chicago, and Philadelphia), results are plotted as the solid lines in Panels 1–4 of [Figure 16.4](#) with distance from the city center on the horizontal axis. [Figure A2](#) in Appendix provides similar plots for the largest 48 metropolitan areas. In all cases, the plots are based on tract-level 2005–2009 pooled ACS data. The patterns in these figures reconfirm that reliance on public transit generally declines sharply with distance from the city center.

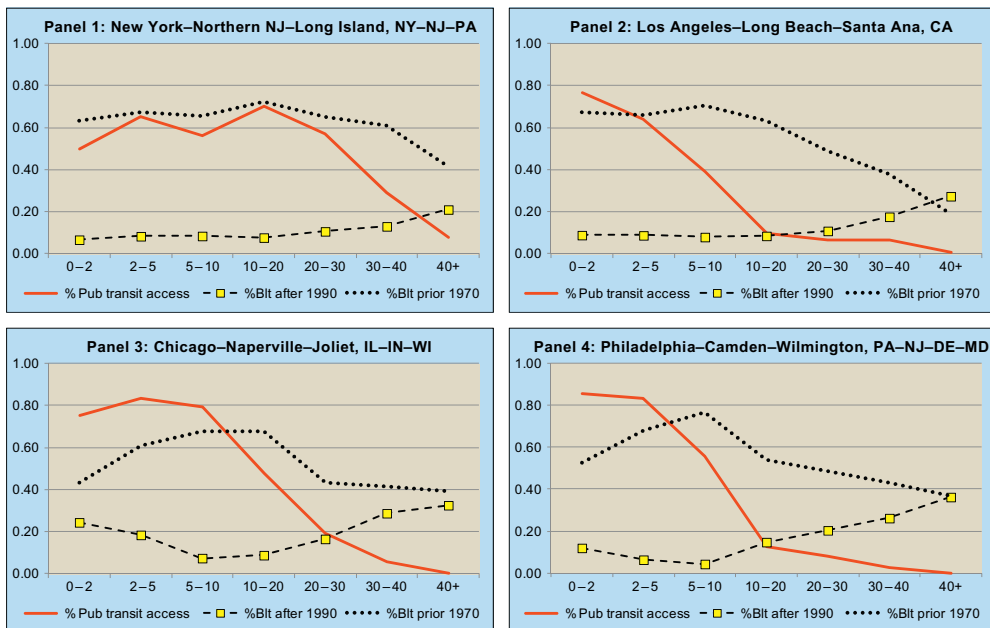


Figure 16.4 Access to public transit and age of the housing stock by distance to the MSA center (2005–2009 ACS pooled sample).

Drawing on these ideas, GKR modified the standard AMM model to allow for the influence of public transit. They first confirmed that car travel is faster than public transit, on average (see table 3, p. 12 of GKR), or $T_P > T_C$, where T_P is travel time per mile by public transit and T_C is travel time by car. If the poor use public transit and the rich drive cars, GKR then showed that the poor will occupy the city center if

$$\varepsilon_{h,y} + \left[\frac{T_P - T_C}{T_P} \left(\frac{Y_{\text{Poor}}}{Y_{\text{Rich}} - Y_{\text{Poor}}} + \varepsilon_{t,y} \right) \right] > \varepsilon_{t,y}, \quad (16.4)$$

where $\varepsilon_{h,y}$ is the income elasticity of demand for land in their discussion. Because the term in brackets is positive, this condition is more likely to be met than $\varepsilon_{h,y} > \varepsilon_{t,y}$, which is the standard model condition for the poor to occupy the city center.¹⁸

A further feature of the model above is that there are fixed costs associated with purchase of autos. Partly for that reason and partly because some rich prefer public transit (as in subway cities), GKR emphasized that alternative parameterizations of the model above can yield a variety of spatial patterns of high- and low-income neighborhoods. As an example, subway travel is faster than car transport in congested areas like Manhattan, downtown Chicago, and Paris. If higher-income families prefer subway to auto travel in these cities, then the model above can readily explain why substantive portions of their downtown areas are high-income. Alternatively, GKR further showed that depending on the level of fixed costs associated with auto purchases, it is possible for a metropolitan area to display alternating bands of high- and low-income neighborhoods as one moves outward from the city center.

16.2.2.1.3 Local public services

As with transportation, the provision of local public services might provide a reason for why the poor live in cities and the rich live in suburbs, even though the rich face a substantially higher time cost for commuting. [Epple and Romer \(1991\)](#), [Ross and Yinger \(1999\)](#), and many other papers demonstrated that with a relatively inelastic supply of housing, households will sort by income across jurisdictions. High-income households will live in jurisdictions with high service levels that low-income households will be unable to access because the rich outbid the poor for housing in such neighborhoods. A limitation of these local public finance models is that they do not incorporate space or commuting to employment centers. For this reason, while these models imply that the poor and the rich are segregated from each other, they say little about where low-income communities are found.

[DeBartolomé and Ross \(2003\)](#) developed a model with an explicit spatial component in which, as in [Glaeser et al. \(2008\)](#), the rich have a higher time cost of commuting. Their

¹⁸ GKR noted that this expression analogous to Equation (16.7) in [LeRoy and Sonstelie \(1983\)](#) who also emphasized the influence of transportation mode and travel speed.

model has two jurisdictions with a central city oriented around a central export node and an outer suburban ring.¹⁹ They proved that stable equilibriums exist in which the poor comprise the majority in the central city and choose a low level of service provision that discourages the rich from locating in the center.

A natural question is what type of processes could lead to this equilibrium? [DeBartolomé and Ross \(2008\)](#) considered this question starting with a model in which all households locate in a single city with an exogenous border. Because of their lower time costs, the poor live closer to the urban fringe and commute further to the city center than the rich. As population grows, two possibilities arise: either the poor move beyond the city boundary and form their own community with low service levels (which are unattractive to the rich) or rising housing prices in the city center cause the rich to form their own community in the suburbs. Simulations in [DeBartolomé and Ross \(2008\)](#) suggest that for a range of realistic parameter values, the rich always develop their own jurisdiction in the suburbs before the poor spill over the city boundary. This leads to the [DeBartolomé and Ross \(2003\)](#) equilibrium in which the poor occupy the center.²⁰

Peer effects have the potential to reinforce the sort of income segregation just described. [Benabou \(1993\)](#), for example, considered an urban model with two neighborhoods where the cost of obtaining skills decreases with the skill of neighborhood residents because of positive peer effects. If households sort across communities based on investment in skills, then peer effects lead to much higher levels of skill inequality across neighborhoods and contribute to stratification of high- and low-income households into different communities.²¹ [Glaeser et al. \(1996\)](#) demonstrated that analogous social interactions can explain the very high variation in crime rates observed across cities. [Calabrese et al. \(2006\)](#) calibrated local public finance models of locational equilibrium to data in Massachusetts with and without peer effects. Adding peer effects to the model strengthens the relationship between jurisdiction income and both local government spending and

¹⁹ See [Epple et al. \(2010\)](#) for an empirical implementation of an income-sorting model with location-specific amenities within political jurisdictions calibrated to data on Pittsburgh. Amenities in the paper implicitly include distance to employment centers and potential commute times. [Hanushek and Yilmaz \(2007\)](#) demonstrated that models of this type generate equilibrium implications that better fit observed sorting patterns.

²⁰ In related work, [Bayer and McMillan \(2012\)](#) simulated a general equilibrium model of residential location and showed that lower commuting costs increase segregation across jurisdictions by income, education, and race.

²¹ In [Benabou's \(1993\)](#) model, individuals choose to be high-skill, low-skill, or out of the labor market. If individuals who choose high skills are able to segregate, the costs of acquiring skills in the lower-skill community increase causing potentially low-skill individuals to opt out of the labor market category. Benabou's analysis of the efficiency implications of peer-based sorting focuses primarily on the labor market outcomes. See [DeBartolomé \(1990\)](#) for an earlier, similar peer effects model, which examines the normative implications for efficiency in the production of skills or education services.

housing prices. This increases the dispersion in jurisdiction quality, which should increase the tendency for income segregation across communities.

More recently, a series of papers have examined the role of peer effects on sorting outcomes associated with education reforms. [Epple and Ferreyra \(2008\)](#) and [Ferreyra \(2009\)](#) calibrated models to match Detroit, Michigan, at a time when an education finance reform had recently been implemented. They found that most of the effect of the reform is on housing prices rather than changes in sorting over demographics or school quality associated with demographic-based peer effects. The key intuition is that the reform did not change the ranking of communities on which sorting is based, and so while changes were capitalized into housing prices, the overall pattern of sorting across jurisdictions remained unchanged.

[Ferreyra \(2007\)](#) used a similar model to examine the effect of a large-scale voucher program in Chicago. In her model, income mixing occurs within neighborhoods because households have heterogeneous preferences for neighborhood locations and neighborhoods differ across multiple dimensions including school quality. She found that school vouchers can reduce income segregation by weakening the link between a family's residential location choice and the quality of schooling opportunities for their children.²²

16.2.2.1.4 Physical amenities

A different explanation for income stratification across communities is that higher-income families are more able and willing to pay for opportunities to live near attractive physical amenities. [Brueckner et al. \(1999\)](#) appealed to this idea in explaining why Paris—the city of light—has a high-income center. They emphasized that Paris has wonderful central city amenities including extensive cultural landmarks, such as the Louvre Museum, the Eiffel Tower, and the Arc de Triumph. [Glaeser et al. \(2008\)](#) also highlighted these features of Paris at the end of their discussion. Amenities also seemingly explain the concentration of high-income households in center-city Vancouver, British Columbia, which sits on a peninsula surrounded by magnificent scenery and other natural amenities. The location of such dramatic historic and/or natural amenities, however, is largely idiosyncratic. For that reason, [Brueckner et al.](#) suggested that amenity-based arguments help explain differences in economic status across communities but do not necessarily imply systematic spatial patterns of where high- and lower-income neighborhoods will be found within a given metropolitan area. In that sense, physical

²² [Epple and Romano \(1998, 2008\)](#) and [McMillan \(2005\)](#) also examined the effect of vouchers in education models with peers, but their models do not allow for sorting across schools located in specific neighborhoods. See [Brunner \(2014\)](#) for a detailed review of the literature on school choice and its effect on sorting and mobility and [Epple and Romano \(2011\)](#) for a more general review of models of education production that involve peer effects.

amenity-based arguments are more of an explanation for “exceptions” to systematic spatial patterns as opposed to characterizing what should be anticipated as the norm.²³

16.2.2.1.5 Nondurable amenities

It should be noted that static amenities need not be physical attributes such as a scenic view or enduring monuments like the Arc de Triumph. [Banzhaf and Walsh \(2008\)](#) examined changes in pollution levels in very small geographic areas between 1990 and 2000. They found that population density increases when exposure to pollution declines and found similar but weaker effects on neighborhood income. They attributed the weaker effects on income sorting to the fact that only large changes in pollution levels will change the relative ranking of neighborhoods and that in equilibrium, it is the ranking of a location that should have the largest effect on income composition, echoing results from [Epple and Ferreyra \(2008\)](#) and [Ferreyra \(2009\)](#).²⁴

It is also worth emphasizing that many models of segregation entail sorting over nondurable amenities (in addition to fiscal services discussed above) that are endogenously determined by the sociodemographic attributes and behavior of whoever lives in the neighborhood. In boundary models like [Courant and Yinger \(1977\)](#) or in tipping models like [Schelling \(1971, 1978\)](#), households sort based on proximity to an individual’s own group including race, ethnicity, or some other distinguishing sociodemographic factor. Building off these models, [Bayer et al. \(2004\)](#) found that sociodemographic characteristics account for much of the segregation by race and ethnicity in the San Francisco Bay Area, while income only plays a limited role in explaining segregation.²⁵ Analogous results are reported in [Bayer et al. \(2014\)](#) who show that black middle-class neighborhoods are more likely to form in US cities where there are a sufficiently large number of higher-income blacks present. In the absence of such a critical mass, middle-income blacks are more likely to either reside in middle-income white communities or forego the higher levels of public services associated with those neighborhoods in favor of locations with greater black representation.²⁶

²³ See also [Lee and Lin \(2013\)](#) for a recent addition to literature related to this theme.

²⁴ [Bayer et al. \(2005\)](#) also documented a strong role of amenities in explaining residential location patterns and further emphasized that amenity values tend to be capitalized into housing price premia.

²⁵ In [Bayer et al. \(2004\)](#), income is the most important driver of black segregation but at most explains only 10% of the observed level of segregation. Over 30% of Hispanic and Asian segregation is attributable to language. Another 20% of Hispanic segregation is attributed to lower levels of educational attainment.

²⁶ [Bayer et al. \(2005\)](#) and [Bayer and McMillan \(2005\)](#) also argued that there are economies of scale in neighborhood formation, and as a result, high-income blacks often face a trade-off between neighborhood amenities (e.g., school quality) and the desired level of African-American presence in a community. In related work, [Bayer and McMillan \(2012\)](#) showed that in the absence of differences in preferences for housing quality, income segregation would be notably reduced because of an increased tendency for high-income blacks to select into lower-income black neighborhoods.

Waldfoegel (2008) provided evidence that endogenous spatial variation in the type of retail services available at the neighborhood level likely contributes to segregation by race and income. Using zip code-level data on restaurants, Waldfoegel (2008) found that fast-food and other lower-priced restaurants are more common in lower-income neighborhoods and that restaurant type differs systematically with the racial and ethnic composition of a community. These patterns are suggestive that restaurants and other local retail establishments cater to the tastes of their local communities. That in turn should reinforce the tendency for individuals to select into neighborhoods populated with others of their own type. To the extent that high- and lower-income families have different tastes for restaurant and other retail services, this mechanism should contribute to income segregation.²⁷

16.2.2.2 *Dynamic models*

The literature described above goes a long way toward helping us understand why central cities tend to be poor relative to the suburbs. Nevertheless, while compelling, the arguments and literature above are inherently static in nature and do not do justice to rising income inequality across neighborhoods and the extensive level of redevelopment and gentrification that have swept through portions of many of the major cities in the United States. The following models help to fill in the pieces of the puzzle.

16.2.2.2.1 *Durable housing, filtering, and gentrification*

Two recent papers, Rosenthal (2008a) and Brueckner and Rosenthal (2009), offer an inherently dynamic view of what drives the spatiotemporal stratification of neighborhoods into high- and lower-income communities. Both papers emphasize that older homes in the United States tend to deteriorate and that higher-income families prefer to live in younger housing, which tends to be in better condition and more attractive. Moreover, because cities tend to develop and redevelop outward over time, the age of homes within a given community tends to be relatively similar. Together, these features imply long-running cycles in which a neighborhood's economic status rises and falls over a period of many decades. This occurs because housing is initially built for higher-income families, decays and is passed down to lower-income families, and is eventually redeveloped and occupied by a new generation of higher-income households.

²⁷ Heterogeneity in preferences for the composition and/or level of local public services could also contribute to segregation. Alesina et al. (1999) is an example of this. Their conceptual model implies that in equilibrium, a neighborhood populated with multiple household types with different preferences for local public services tends to adopt lower public service levels. Empirical evidence in their paper confirms that ethnic fragmentation is negatively related to local spending on welfare support. An implication is that ethnic groups may sort into segregated neighborhoods in order to secure their preferred public service bundles.

Brueckner and Rosenthal (2009) and Brueckner (2011, chapter 3) highlighted implications of the correlation between the timing and location of development/redevelopment. In growing cities, housing at the urban/rural fringe tends to be newly developed and occupied by higher-income families. Among older metropolitan areas, newly built housing will also be found in one or more rings at different distances from the center where redevelopment has recently occurred.²⁸ Such locations are also populated with higher-income families who seek out the newer housing. Moreover, as redevelopment bands gradually move outward from the city center, the locations of high- and low-income neighborhoods shift, and this causes the relative economic status of the central cities and the suburbs to cycle up and down over extended periods of time.

Figure 16.5 displays the sort of housing stock age profile that might be anticipated in a growing city given the slow decay of housing capital and periodic redevelopment. Evidence in support of the qualitative features of the durable-housing model is provided in Figure 16.4. In the four panels described earlier (when discussing public transit access), plots are also provided of the frequency of new and old housing stock at different distance bands as one moves away from the city center. As before, the plots are based on census tract data from the pooled 2005–2009 ACS. Similar plots are provided for all 48 of the largest US MSAs in Figure A2 of Appendix.

Consider Panels 3 and 4 of Figure 16.4 first, for Chicago and Philadelphia, respectively. For both MSAs though especially for Chicago, it is evident that close to the city center, there has been considerable new construction of residential units as seen in the dashed-marked lines that plot the percent of housing stock built after 1990. Also apparent, the share of housing stock built prior to 1970 displays a marked upward trend as one moves away from the city center (as seen in the dotted lines). In both metro areas, the concentration of older housing stock peaks roughly 10–20 miles from the center after which the incidence of newly built housing increasingly dominates. These patterns mirror the dramatic v-shaped plots for neighborhood economic status in Chicago and Philadelphia as displayed in Figure 16.1. Together, these patterns point to extensive recent redevelopment and gentrification that has occurred in areas close to the center of these metropolitan areas.²⁹

In contrast, for New York and Los Angeles (Panels 1 and 2), the plots suggest that central city redevelopment has been less pronounced than in Chicago and Philadelphia, allowing for exceptions such as Harlem and other such areas. Observe, for example, the

²⁸ See Braid (2001) for an earlier model in which the optimal redevelopment location moves outward over time and the metropolitan age distribution of housing is nonmonotonic over space. Notably in Braid (2001), developers are forward-looking and consider the time path of population growth in redevelopment decisions. See also Arnott and Braid (1997) for a model of filtering with both redevelopment and maintenance.

²⁹ Analogous patterns are also present in Boston, Miami, Minneapolis–St. Paul, Seattle, Baltimore, Denver, and Portland Oregon, as shown in Figures A1 and A2 of Appendix.

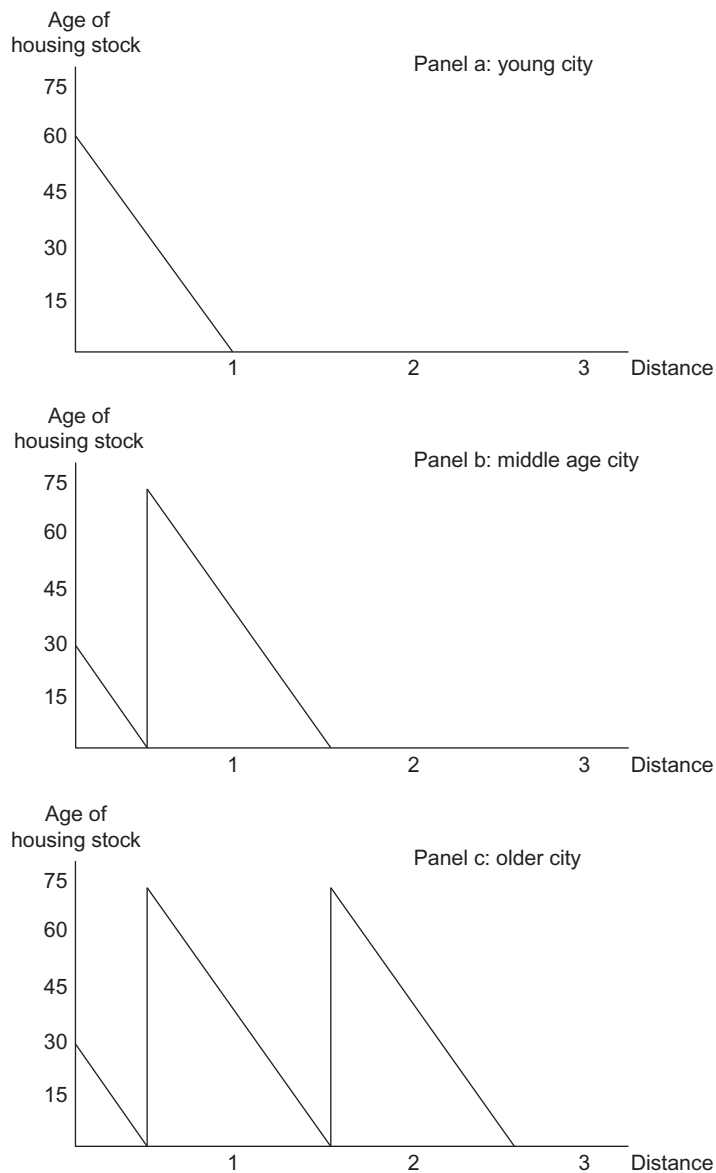


Figure 16.5 Aging housing stock and income stratification. Panel (a) young city. Panel (b) middle age city. Panel (c) older city.

greater relative presence of older homes (the dotted lines) closer to the city centers. These patterns are also echoed in the plots of neighborhood economic status in [Figure 16.1](#): for New York, economic status varies little on average from the city center out to 20 miles after which it rises, while for Los Angeles, economic status increases monotonically with distance from the center.

16.2.2.2.2 Social dynamics, segregation and tipping, and homeownership

While [Rosenthal \(2008a\)](#) found strong evidence of mean reversion in neighborhood income, he also found that sociodemographic variables have substantial power in explaining the rate at which neighborhoods transition up and down the income distribution even after conditioning on the age distribution of the housing stock, access to public housing, and other factors. Neighborhoods with high levels of positive demographic attributes, such as resident education levels, are much more likely to hold their position in the income distribution as the housing stock ages. Accordingly, in this section, we review a set of papers that shed light on the role of social dynamics in driving change in neighborhood economic status. A theme throughout is that while these forces are important drivers of neighborhood change, as with physical amenities, they do not necessarily imply a systematic spatial pattern that would allow one to anticipate where high- and low-income communities will be found. A further distinction is that social dynamics tend to be self-reinforcing in a manner that contributes to the endogenous formation of high- and low-income communities.

Two recent papers, [O'Sullivan \(2005\)](#) and [Guerrieri et al. \(2013\)](#), are examples of this. Both papers treat the presence of higher-income families as itself an attractive force that draws additional high-income families to a community. In [O'Sullivan \(2005\)](#), the presence of high-income families is assumed to lower local crime rates, while high-income families are further assumed to have a higher willingness to pay for safe communities. These two assumptions ensure that the arrival of a set of high-income families has a self-reinforcing effect that attracts additional higher-income households causing a neighborhood to gentrify. O'Sullivan offered descriptive evidence that Portland, Oregon, in the 1990s underwent just such a transformation. More generally, any change in neighborhood income will tend to be self-reinforcing if higher levels of neighborhood income contribute to improved levels of local amenities, as with school quality, local aesthetic features including gardens, trimmed lawns and clean streets, or improved word-of-mouth job networks.³⁰

[Guerrieri et al. \(2013\)](#), hereafter GHH, took a more direct approach to these ideas by assuming that families directly value proximity to higher-income households. This generates many of the same equilibrium outcomes as in [O'Sullivan \(2005\)](#). GHH further argued that following a city-wide positive demand shock, the growing population of high-income households will increasingly seek housing in lower-income communities adjacent to existing high-income neighborhoods, thereby expanding the geographic extent of existing high-income areas. This should also have the effect of raising house prices more in low-income communities that abut high-income enclaves, contributing

³⁰ [Bond and Coulson \(1989\)](#) developed a model that incorporates such income externalities into a model with aging housing stocks and filtering. They argued that departures of higher-income families as homes age can lead to rapid downward transition of a neighborhood.

to within-city variation in house price growth. GHH tested their model using quarterly Case–Shiller zip code-level repeat sales house price indexes from the late 1980s to 2008 along with census tract data from the 1980, 1990, and 2000 census. Results support the model predictions.³¹

The above models of income segregation build on the long literature on racial segregation in housing including Schelling's (1971, 1978) well-known tipping models. Schelling's models are sometimes classified into two types: the spatial proximity model and the isolated neighborhood model. In the spatial proximity model, households are arrayed along a line (or a lattice) with mobile individuals selecting their optimal position on the line. Schelling (1971, 1978) assumed a simple preference function in which utility differs based on a step function over local racial composition. Households receive high levels of utility when the share of a household's own group in the immediate neighborhood is above a fixed threshold and low levels of utility when household group share is below the threshold. Households take turns moving along the line and choose location to maximize utility treating the current spatial distribution of household types as given. The classic result in these models is that even with very small differences in preferences for neighborhood racial composition, perfect segregation is the equilibrium outcome.

Recent studies (see Zhang, 2004, 2011; Pans and Vriend, 2007; O'Sullivan, 2009) have shown that these results are robust to more complicated preference functions. Even when individuals have relatively strong preferences for integration, for example, a myopic, best response allocation process of individuals across residential locations leads to segregation provided there is at least some difference across household types in preference for integration (e.g., whites prefer a slightly less integrated neighborhood than minorities). Pans and Vriend (2007), for example, considered a model in which the utility maximizing level of integration is the same for two groups of households. A key feature of their model is that the rate at which utility declines with departures from the optimum level of integration is faster as own-group share falls below the optimum as compared to when own-group share rises above the optimum. This sort of asymmetry of preferences leads to perfect segregation even though the set of Nash equilibrium does not include perfect segregation. A fundamental feature of all of these models is that integrated neighborhoods tend to be unstable and prone to tipping toward a segregated outcome.

The intuition behind this result is relatively straightforward. Asymmetrical preferences as just described will cause individuals belonging to a minority group to favor integrated communities with relatively high shares of minority households over integrated communities with lower minority shares. As minority households choose away from

³¹ It is worth noting that the Guerrieri et al. (2013) model implies a particular spatial form of gentrification in which existing higher-income communities tend to grow into adjacent lower-income areas. This is in contrast to dynamic models such as Brueckner and Rosenthal (2009) where the spatiotemporal pattern of newly built housing drives the location of high-income communities.

low minority-share integrated neighborhoods to high minority-share integrated neighborhoods, integrated neighborhoods become increasingly segregated toward either minority or majority households.

In the isolated neighborhood model, [Schelling \(1971, 1978\)](#) still assumed that household preferences for neighborhood racial composition are based on a step function. However, in this case, majority and minority households have heterogeneous preferences for integration such that the critical own-group neighborhood share (or step) at which utility changes is lower for more tolerant individuals and higher for individuals with a stronger taste for segregation. In this setup, integrated neighborhoods contain the most tolerant members of each group. Moreover, an increase in one group's representation within a neighborhood reduces the appeal of the neighborhood to the alternate group. This type of model typically yields a minimum of three equilibriums for a single neighborhood: all-majority (e.g., white in the United States.), all-minority, and integrated where the marginal-preference individual in each group is just indifferent between the integrated neighborhood and an outside option. Importantly, the middle, integrated neighborhood is a knife-edge equilibrium and unstable. This is because a very small increase in one group's population will cause the marginal individuals that belong to the alternate group to move away causing group composition within the neighborhood to shift toward an all-majority or all-minority equilibrium. As with the spatial proximity model, this implies that integrated neighborhoods are prone to tipping and tend to be in transition toward segregated outcomes.

[Card et al. \(2008, 2010\)](#), hereafter CMR, introduced a new model of tipping building on the isolated neighborhood model of Schelling. A distinguishing result from their model is that when neighborhoods tip, they do so toward an all-minority composition but not toward an all-majority composition. They refer to this as one-sided tipping.

In the CMR modeling framework, stable integrated neighborhoods emerge when whites (the majority group in their discussion) have a sufficiently strong taste for living in integrated communities. If, however, minority demand for housing in the integrated community rises above a critical value, the minority representation in the integrated neighborhood will become too high to be supported by white preferences for integration, and as a result, whites flee the neighborhood and the community evolves to an all-minority, segregated outcome. The stronger white household tastes are for integration, the higher the tipping point defined by neighborhood minority share. In this context, tipping can be prompted by a change in minority housing preferences (as driven by an increase in income, for example) and assuming that housing quality is higher in integrated neighborhoods as compared to all-minority communities.

Using a panel of 1970–2000 census tracts from the United States, [CMR \(2008\)](#) found strong evidence of tipping behavior in most cities and suburbs with tipping points between 5% and 20% minority household share (relative to majority white households). They also found that tipping points are lower in metropolitan areas where whites have

more tolerant attitudes. Importantly, tipping appears to be one-sided as neighborhoods with minority shares above the tipping point for their metro area exhibit rising minority shares. In contrast, neighborhoods with minority shares below their tipping points exhibit relatively stable minority shares.

These sorts of tipping models make clear that when households choose neighborhoods in part based on the attributes of their neighbors (e.g., race, ethnicity, or income), small perturbations in migration patterns or preferences can destabilize integrated neighborhoods and lead to segregated communities. It is less clear, however, whether tipping will be one-sided or two-sided when households sort based on the income of their potential neighbors as this depends on the strength and nature of preferences for mixed-income communities.

A very different set of social interactions tend to be associated with homeownership and can have strong stabilizing effects on neighborhoods. Homeowners are financially invested in their communities in a manner that is not really matched by renters. Homeowners, therefore, have financial incentives to behave in ways that enhance local property values, whether by performing more maintenance on both the interior and exterior of their homes, gardening, participating in neighborhood watch groups, or voting for local policies that will increase neighborhood appeal. Homeowners also tend to be less mobile than renters. These differences suggest that the presence of homeowners may reduce any tendency for neighborhood economic decline and thereby strengthen the current and future economic status of a community.³²

Consistent with these ideas, using 1970–2000 census tract data for the United States, [Rosenthal \(2008a\)](#) found that higher homeownership rates reduce the likelihood that a neighborhood transitions down in economic status even after controlling for income, education, racial composition, and a host of other socioeconomic attributes of a community. Moreover, because homeownership is more prevalent in higher-income neighborhoods, its presence helps to explain the greater tendency for higher-income communities to retain their current economic status as shown in [Figure 16.2](#) and discussed earlier.³³

An increasingly rich literature has provided evidence of the many ways in which homeownership likely contributes to neighborhood stability. Important among these, the homevoter hypothesis ([Fischel, 2001](#)) suggests that homeowners are more likely to vote for local policies that enhance neighborhood quality and raise property values. Such behavior will tend to reinforce sorting outcomes described above and strengthen a community's economic status. [Brunner et al. \(2001\)](#) using aggregate data and [Brunner and Sonstelie \(2003\)](#) using micro-based survey data, for example, provided

³² See [Ross \(2011\)](#) for a short survey of this literature.

³³ The strong positive correlation between homeownership, income, and education is overwhelming and has been well documented (e.g., [Boehm and Schlottmann, 2004](#); [Haurin et al., 2007](#)).

evidence that homeowners in California are more likely to vote against school vouchers in jurisdictions where passage of the voucher referenda would reduce property values. Hilber and Mayer (2009) showed that school spending increases when the land available for new development is limited, an implication of which is that housing supply is inelastic and that rising school quality will be capitalized into higher housing prices. Ahlfeldt and Maennig (2014) report that positive price effects of a proposed airport development plan in Berlin prompt greater voter support for the plan in locations populated with higher concentrations of homeowners. From this, they infer that homeowners are more responsive than renters to positive capitalization effects, consistent with the home voter hypothesis.^{34,35}

A very different perspective on the influence of homeownership on neighborhood economic status and change is offered by Boustan and Margo (2013). They provided evidence that white flight between 1940 and 1980 substantially increased black homeownership rates in US central cities (they instrumented for white flight using Baum-Snow's (2007) planned highway network measures). Boustan and Margo argued that this occurred in part because white flight contributed to lower central city home prices, which would have facilitated purchase of homes by low-wealth black families. Deng et al. (2003) and Dawkins (2005) also provided evidence that lower home prices contribute to minority homeownership. These studies point to a subtle but important implication of homeownership and the distribution of wealth for neighborhood income stratification and related dynamics. On the one hand, homeownership enhances a community's economic status and raises property values, but lower property values facilitate homeownership among low-wealth homebuyers.³⁶ These arguments suggest that homeownership has both direct and indirect effects on a community's future economic status by affecting the type of families that have sufficient wealth to access the neighborhood.

³⁴ Dehring et al. (2008) provided analogous evidence when considering the impact of a 2004 referendum on the development of a publically subsidized football stadium for the Dallas Cowboys in Arlington, Texas. As with Ahlfeldt and Maennig (2014), they first use hedonic methods to document spatial patterns of post-announcement but prereferendum effects on local home prices. In a second stage, voting patterns are assessed in part as a function of anticipated changes in home prices. Dehring et al. (2008) reported positive effects of home price capitalization on support for the stadium but did not find differential effects of home price capitalization-based precinct-level variation in homeownership rates (see Dehring et al., 2008, p. 164). This latter result differs from Ahlfeldt and Maennig and is suggestive of the absence of a home-voter effect.

³⁵ Homeowning residents may also support zoning ordinances that limit the density of development in an attempt exclude lower-income residents and maintain a community's economic status. See Ross and Yinger (1999) for a discussion on the use of zoning to influence a jurisdiction's income and economic status.

³⁶ It is well appreciated that household wealth is a critical driver of homeownership (see, for example, Charles and Hurst, 2002; Gabriel and Rosenthal, 2005; Haurin et al., 2007; Ross and Yinger, 2002, chapter 2, to name just a few references).

16.2.3 Evidence of drivers of neighborhood economic status

From the discussion above, it is clear that several factors drive static and dynamic spatial patterns of neighborhood economic status. The first of these is the standard AMM model that highlights tension between the time cost of commuting and the income elasticity of demand for housing. Properly parameterized, this model suggests that economic status is likely to decline with distance from employment centers as emphasized by Glaeser et al. (2008). The second mechanism is access to public transit that suggests the opposite that the poor occupy the center. The third is the role of local public services, which tend to be of higher quality (e.g., school quality) in higher-income communities, thereby reinforcing the spatial implications of other mechanisms that are at work. The fourth is the slow deterioration of existing housing stock over time coupled with the correlation between the timing and location of development and redevelopment of housing stock. A fifth mechanism is the role of location-specific amenities that need not display a systematic spatial pattern but that nevertheless attract or deter households. Such amenities include physical features of an area like scenic views and proximity to a lakefront, as well as amenities that arise endogenously like the racial/ethnic composition of a community, ethnic restaurants, and income itself.

Given the different implications of these mechanisms, it is not surprising that there is considerable variation in spatial patterns of neighborhood economic status among the largest 48 MSAs as shown in Figure A1 in Appendix.³⁷ This is especially evident from the sharp differences in spatial patterns of neighborhood economic status in New York, Los Angeles, Chicago, and Philadelphia in Figure 16.1 as discussed earlier. Overall, distance to the MSA center alone is a rather weak predictor of a neighborhood's economic status.

Brueckner and Rosenthal (2009) and Glaeser et al. (2008) both run a series of regressions that help to confirm the effect of each mechanism discussed above on spatial variation in neighborhood economic status. Select results from the two papers are reproduced in Table 16.2 where estimates in columns 1 and 2 are drawn from Brueckner and Rosenthal (2009) and estimates in columns 3 and 4 are drawn from Glaeser et al. (2008). It is worth noting that the dependent variable in columns 1 and 2 from Brueckner and Rosenthal (2009) is neighborhood relative income in 2000 calculated as in Figure 16.1 (and Figure A1 in Appendix). The dependent variable in columns 3 and 4 from GKR is log of census tract median income in 2000.

The first column in Table 16.2 includes only distance from the center as a control along with MSA fixed effects that capture time-invariant MSA-level amenities and

³⁷ It is worth noting that while urban poverty is especially concentrated in the central cities (e.g., Glaeser et al., 2008; Rosenthal, 2008b), there is considerably more spatial variation in neighborhood median income.

Table 16.2 Year 2000 neighborhood economic status in [Glaeser et al. \(2008\)](#) and [Brueckner and Rosenthal \(2009\)](#)

	Track Avg Inc./MSA Avg Inc. ^a		Log (median income) ^b	
	Simple model	With controls	Tracts without public transit	All tracts
Miles from center	0.0018 (9.08)	−0.0010 (−4.28)	— —	— —
Miles from center * within 3 miles	—	—	—	0.2214 (0.0044)
Miles from center * more than 3 miles	—	—	—	0.0513 (0.0013)
Miles from center * within 3 miles * subway city	—	—	—	−0.3523 (0.0129)
Miles from center * more than 3 miles * subway city	—	—	—	−0.0039 (0.0032)
Miles from center * 5–10 miles	—	—	−0.0017 (0.0044)	—
Miles from center * 10–15 miles	—	—	−0.0219 (0.0055)	—
Access to public transit in 1980	—	−0.0945 (−10.76)	—	—
% Dwellings 0–4 years in 1980	—	0.9967 (34.01)	—	—
% Dwellings 5–9 years in 1980	—	0.3219 (9.98)	—	—
% Dwellings 10–19 years in 1980	—	0.6197 (21.49)	—	—
% Dwellings 20–29 years in 1980	—	0.6501 (18.22)	—	—
% Dwellings 40 or more years in 1980	—	0.5365 (16.66)	—	—
Constant	0.9772 (300.09)	0.4765 (18.53)	10.7011 (0.0400)	9.9893 (.0103)
Observations	50,511	48,437	1394	27,218
MSA fixed effects	Yes	—	Yes	Yes
Unified school district fixed effects	—	Yes	—	—
Adj. <i>R</i> -square	0.0016	0.1747	0.437	0.342

^aReproduced from [Brueckner and Rosenthal \(2009\)](#), table 1, p. 733). Public-transit access equals 1 if 10% or more of the tract's population in 1980 used public transit and 0 otherwise. Dwelling ages are measured as of 1980. The omitted category is age 30–39 years.

^bReproduced from [Glaeser et al. \(2008\)](#), table 6, p. 17.

related attributes. The familiar pattern is obtained with economic status rising with distance from the center.

Column 2 replaces MSA fixed effects with school district fixed effects that control for local-level public services and amenities. Additional controls are also added for public transit access and the age distribution of the housing stock, all measured at the census tract level in 1980, 20 years prior to the dependent variable. These controls address either directly or indirectly all of the drivers of neighborhood economic status discussed above.

Results in column 2 confirm that access to public transit has a powerful attractive effect on lower-income families. This is evident from the highlighted significant and negative coefficient on public transit. The magnitude of that coefficient implies that, all else equal, communities with good access to public transit have median income relative to the MSA that is 9.45 percentage points lower than communities for which public transit is not accessible.

Interpretation of the dwelling age variables is more subtle. Note first that the dwelling age distribution is broken into six categories based on the percentage of homes in the census tract that are 0–4, 5–9, 10–19, 20–29, 30–39, and 40 or more years in age. The omitted dwelling age category is the percent of homes between ages 30 and 40 years, and all of the dwelling age variables are measured using 1980 data, 20 years prior to the dependent variable. The positive and significant coefficients on all of the included dwelling age categories imply a u-shaped pattern with respect to the influence of the 20-year lagged age distribution of the housing stock on current neighborhood economic status. That pattern is consistent with a scenario in which homes tend to depreciate over time, on average, and that older housing is eventually demolished and replaced with new dwellings. This is because few homes aged 30–40 years in 1980 would have been demolished by 2000 with demolitions instead concentrated in homes that were over 40 years in age in 1980. The positive coefficient on homes over 40 years in age is therefore suggestive that older housing stocks are being replaced, which draws in higher-income families. The positive coefficient on dwelling age segments under 30 years in age is suggestive that such homes offer more valuable service flows in 2000 than homes that were 30–40 years in age in 1980.

Having controlled for the influence of public transit, local amenities, and the age of the housing stock, in principle, the AMM model prediction should emerge that income falls with distance from the center. This is precisely what is obtained as indicated by the negative and highly significant coefficient on distance from the center.³⁸ The magnitude

³⁸ Brueckner and Rosenthal (2009) reported two additional intermediate models with specifications sandwiched between those reproduced in columns 1 and 2 of Table 16.2. Replacing MSA fixed effects with school district fixed effects reduces the distance coefficient from 0.0018 (with a *t*-ratio of 9.08) to 0.00005 (with a *t*-ratio of 0.23). Adding the control for public transit changes the distance coefficient to 0.00011 (with a *t*-ratio of 0.45) while yielding a coefficient on public transit of -0.1525 (with a *t*-ratio of 18.06). Only when the age distribution of the housing stock is added to the regression—as reproduced in column 2 of Table 16.2—does the distance coefficient become negative and significant.

of the estimated coefficient indicates that tract income relative to MSA income falls roughly 1 percentage point with every 10 miles distance from the center.

Column 3 reports results from a restricted sample in [GKR \(2008\)](#) designed to control for the same confounding factors above that might otherwise mask the anticipated AMM pattern. Specifically, the regression is based on a sample of census tracts 5–15 miles from the city center drawn from 99 select MSAs for which public transit access is all but absent within this distance band. GKR referred to these census tracts as “car zone” areas. They emphasized that since public transit is essentially not available in these areas, households have no viable option except to rely on automobiles to travel to and from work (and for nonwork trips). Since all households rely on car travel, the influence of public transit on income sorting disappears. Although not discussed in GKR, it seems likely that these 99 metropolitan areas are smaller MSAs, and for that reason, housing stock between 5 and 15 miles from the center will tend to be relatively similar in age. Their sampling strategy, therefore, likely also reduces the influence of the dwelling age on income sorting. The model also includes MSA fixed effects that help to control for MSA-level amenity effects.

The [GKR \(2008\)](#) model also recovers the AMM model prediction that neighborhood income should fall with distance from the center. As reported in column 3, with each mile from the center, median income in the census tract declines roughly 1.7% in the 5–10 mile distance band and 2% in the 10–15 mile band. Both estimates are also highly significant.

In summary, evidence from [Glaeser et al. \(2008\)](#) and [Brueckner and Rosenthal \(2009\)](#) confirms that public transit and aging housing stocks both attract lower-income families, reducing a neighborhood’s income and economic status, on average. Controlling for these factors and local amenities (through location fixed effects), both studies also provide evidence that supports a central prediction of the AMM model: neighborhood economic status falls with distance from the city center as would be expected given compelling evidence that the income elasticity of demand for housing is smaller than the income elasticity with respect to the time cost of commuting. Two further implications follow. First, it should come as no surprise that spatial patterns of income sorting differ across US metropolitan areas even while the norm is for poor cities and rich suburbs. Second, the sorting mechanisms emphasized here include static drivers such as public transit access and dynamic drivers such as the slow deterioration of existing housing stock.

16.2.4 The pace and timing of change in neighborhood economic status

The discussion above makes clear that for many communities, neighborhood economic status is dynamic rather than fixed. This section considers the pace at which communities rise and fall and sheds further light on drivers of the rate of change. We begin by discussing literature on home maintenance that affects the rate at which housing deteriorates over time. This leads naturally into discussions of the timing of demolition and

redevelopment decisions and the rate at which neighborhoods and individual homes filter down and up the economic ladder.

16.2.4.1 Home maintenance and depreciation of housing capital

Home maintenance is an essential but understudied feature of housing supply that has a direct impact on the rate at which housing depreciates. Any homeowner who has replaced a leaky roof, installed a new furnace, or repaired rotted window frames knows this at an intimate level. Any discussion of the pace at which neighborhoods rise and fall must therefore recognize that maintenance decisions play a role.

Table 16.3 reported summary measures from the US Bureau of Economic Analysis (BEA) that highlight the magnitude of investments in home maintenance. Values are provided for every other year from 2000 to 2012 for nominal expenditures on new construction and maintenance. Several patterns are especially relevant to our discussion of neighborhood dynamics.

Observe that maintenance expenditures are large, both in level terms, as compared with expenditures on new home construction and as compared to GDP. This is evident in the first four rows that report expenditures on new construction and maintenance in dollar values and also as a percentage of GDP. Expenditures on new home construction are roughly 2.5% of GDP prior to the 2004–2006 housing boom, peak at roughly 3.4% of GDP in 2004–2006, and fall back to just under 1% of GDP in 2010. During this same period, expenditures on home maintenance and improvements ranged between roughly 1% and 1.3% of GDP.

Also apparent in the middle rows of the table, roughly 90% of expenditures on new construction are for owner-occupied rather than rental units. It is well documented that owner-occupiers are of higher-income status than renters. This pattern therefore echoes our earlier discussion that new construction disproportionately accommodates higher-income households and is associated with higher-income neighborhoods.

A final pattern in the table worth highlighting is that roughly three-quarters of expenditures on home maintenance and improvements are in the owner-occupied sector. This is notably less than the owner-occupied share of new construction but still substantially above the overall US homeownership rate, which ranged between 65% and 69% over the 2000–2012 period. In part, these patterns arise because as homes age, there is a net tendency for units to switch from owner-occupied to rental occupancy (see [Rosenthal, 2014](#)). We consider now the degree to which home maintenance slows depreciation of housing stock.

[Harding et al. \(2007\)](#) used 1985–2001 data from the AHS to investigate the impact of home maintenance expenditures on the rate of house price inflation (quality adjusted) among owner-occupied homes. They documented that reported home improvements and maintenance average just over 1% of house value on an annual basis, consistent with the large expenditure on home maintenance and improvements documented in

Table 16.3 Investment in US residential fixed assets (excludes manufactured housing)^a

	2000	2002	2004	2006	2008	2010	2012
New construction (million \$) ^b	263,657	295,481	414,619	466,193	225,024	122,298	151,347
Percent of GDP	2.56%	2.69%	3.38%	3.36%	1.53%	0.82%	0.93%
Repairs/improvements (million \$) ^c	116,725	133,974	159,134	183,626	170,218	159,609	159,505
Percent of GDP	1.13%	1.22%	1.30%	1.33%	1.16%	1.07%	0.98%
New construction—owner-occupied (million \$)	238,055	266,519	376,838	418,887	192,574	109,545	132,891
Percent of total new construction ^d	90.29%	90.20%	90.89%	89.85%	85.58%	89.57%	87.81%
New construction—rental (million \$)	25,602	28,962	37,781	47,306	32,450	12,753	18,456
Percent of total new construction ^d	9.71%	9.80%	9.11%	10.15%	14.42%	10.43%	12.19%
Repairs/improvements—owners (million \$)	81,050	98,759	117,782	146,459	132,094	124,218	123,576
Percent of total repairs and improvements ^d	69.44%	73.72%	74.01%	79.76%	77.60%	77.83%	77.47%
Repairs/improvements—rental (million \$)	35,675	35,215	41,352	37,167	38,124	35,391	35,929
Percent of total repairs and improvements ^d	30.56%	26.28%	25.99%	20.24%	22.40%	22.17%	22.53%
Gross domestic product (billion \$) ^e	10,290	10,980	12,277	13,858	14,720	14,958	16,245

^aSource: US Bureau of Economic Analysis (BEA), Detailed Data for Fixed Assets and Consumer Durable Goods, Residential Detailed Estimates; URL: <http://www.bea.gov/national/FA2004/Details/xls/detailresidential.xlsx>.

^bNew investment in residential units, not including manufactured homes or acquisition and disposal costs.

^cRepairs and improvements computed as the sum of three components: additions and alterations, major replacements, and equipment (for tenant-occupied only).

^dTotal new construction and total repairs and improvements based on the sum from owner-occupied and rental units.

^eSource: BEA, Gross Domestic Product, NIPA tables, section 1; http://www.bea.gov/national/nipaweb/GetCSV.asp?GetWhat=SS_Data/Section1All_xls.xls&Section=2.

Table 16.3. Harding et al. (2007) then augmented a standard repeat sales model (e.g., Case and Shiller, 1987) to include controls for age-related depreciation and maintenance expenditures in order to estimate the impact of these additional controls on house price depreciation rates. Their model takes on the following form:

$$\ln \left(\frac{P_{t+\tau,i}}{P_{t,i}} \right) = \sum_{i=1}^{\tau_i} \gamma_i D_{t,i} + \alpha \log(\tau_i) + \tilde{b}_1 \frac{\tilde{M}_{1-3,i}}{P_{t,i}} + \tilde{b}_3 \frac{\tilde{M}_{4-6,i}}{P_{t,i}} + \tilde{b}_3 \frac{\tilde{M}_{7-18,i}}{P_{t,i}} + \omega_{t+\tau,i}, \quad (16.5)$$

where P_t and $P_{t+\tau}$ are a home's sale price upon turning over at time t and $t+\tau$. The standard repeat sales model includes only $\sum_{i=1}^{\tau_i} \gamma_i D_{t,i}$ for control measures with $D=1, 0$, or -1 for the second sale date $t+\tau$, dates for which the home does not sell, and the first sale date t . That specification is obtained by differencing hedonic house price regressions between the two sale dates and imposing the assumption that all attributes of the home and their shadow prices remain fixed over time. In the augmented model in expression (16.5), depreciation associated with the passage of time is captured by $\alpha \log(\tau)$, while the $\tilde{M}$ terms capture the level of maintenance relative to period- t house value 1–3 years in the past, 4–6 years in the past, and 7–18 years in the past. A limitation of the model is that depreciation must be entered nonlinearly (as $\log(\tau)$ in this instance) to avoid perfect collinearity with the inflation indexes $\sum_{i=1}^{\tau_i} \gamma_i D_{t,i}$. The gross-of-maintenance rate at which homes depreciate is then given by α/τ .

Evaluating using sample median values for τ , Harding, Rosenthal, and Sirmans reported that the average annual gross-of-maintenance real rate of house price depreciation is roughly 3%. Expenditures on maintenance offset roughly 1 percentage point of that depreciation reducing the net-of-maintenance real rate at which homes depreciate to 1.94% per year (see table 4, p. 212 of Harding et al., 2007). Together, these results and the summary measures in Table 16.3 confirm that expenditures on home maintenance are large and substantially slow the rate at which housing depreciates.

Gyourko and Saiz (2004) showed that homeowners reduce maintenance expenditures by up to 50% if their home falls in value below construction cost, a result that has important implications for neighborhood dynamics.³⁹ Consider, for example, a city in which home values have fallen below construction cost as would be typical of many declining areas such as Detroit that have experienced sharp declines in employment. In the face of declining house prices, the return on home maintenance will be far less than 1\$ for every dollar invested. Moreover, in locations that have been subject to sharply falling home prices, many homeowners find themselves in a negative net equity position in which they

³⁹ As with Harding et al. (2007), they obtained this result using the AHS panel.

owe their mortgage lender more than the market value of the home. Such families have financial incentives to default on their mortgage upon moving from their homes and, for that reason, have little incentive to maintain their home.

These implications of [Gyourko and Saiz \(2004\)](#) suggest that declining communities are likely to receive less investment in home maintenance. Reduced home maintenance in turn will accelerate depreciation of the housing stock and contribute to neighborhood (and city-level) decline. [Brueckner and Helsley \(2011\)](#) developed a dynamic model that reinforces this view. They argued that market failures that draw households to suburban communities in search of more attractive homes depress central city home prices and undermine incentives to maintain central city housing. This contributes to decay of the central city housing stock and leads to further central city decline. Together, [Gyourko and Saiz \(2004\)](#) and [Brueckner and Helsley \(2011\)](#) suggested that there are important dynamic links between maintenance and neighborhood change although that relationship has mostly escaped attention in the literature.⁴⁰

16.2.4.2 *The decision to redevelop*

The model in [Brueckner and Rosenthal \(2009\)](#) discussed earlier emphasizes that as housing ages, it slowly deteriorates and is eventually demolished and redeveloped. As described above, this contributes to systematic cycles in neighborhood economic status as higher-income families are attracted to the newly built housing. This section considers a set of papers that evaluate the timing and nature of the decision to redevelop a home.

[Rosenthal and Helsley \(1994\)](#) developed and test a model that frames the decision to redevelop a parcel of land. In a perfect foresight setting, they argued that housing is redeveloped when the value of land in its vacant state exceeds the value of the property with the existing structure in place and adjusted for demolition costs. Empirical evidence based on single-family detached home sales in 1987 in Vancouver, British Columbia, supports the model structure.⁴¹

[Rosenthal and Helsley \(1994\)](#) separated home sales into two samples, those sold for redevelopment and those for which the existing structure is retained. Because demolition costs are small for single-family detached homes, they argued that properties sold as tear-downs can be treated as vacant land sales. Allowing for sample selection effects, sale price regressions are then estimated using the two samples. A second-stage structural probit model yields support for the view that homes are demolished when the site is more

⁴⁰ [Lee et al. \(2013\)](#) offered a completely different perspective on home maintenance. They argued that with incomplete information about house quality, home maintenance may be perceived by prospective buyers as a signal that housing quality is low. Under this argument, home maintenance reduces house value, *ceteris paribus*. An implication of this argument is that uncertainty about housing quality could accelerate deterioration of the stock by reducing incentives for maintenance.

⁴¹ Vancouver is a particularly convenient location for such a study as it underwent a wave of redevelopment during this period with nearly 8% of single-family homes sales in 1987 having been sold for redevelopment.

valuable as vacant land, which allows for future development options as compared to retaining the current structure and use.

Dye and McMillen (2007) revisited the redevelopment question using data from seven municipalities in the Chicago metropolitan area. Between 1996 and 2003, the share of existing stock that was redeveloped ranged from a low of 2.11% (in Park Ridge) to a high of 9.4% (in Winnetka). The city of Chicago experienced a redevelopment rate of 2.9%. These high rates of redevelopment are consistent with the patterns in Figure 16.1, Panel 3, and Figure 16.4, Panel 3. As discussed earlier, those panels point to extensive redevelopment and gentrification in Chicago in recent years. Using these data, Dye and McMillen also found compelling support for the principle that the price of teardown properties is approximately equal to the value of vacant land.⁴²

Three recent papers more clearly cast the decision to redevelop or partially modify a structure as a call option held by the owner of the property. That option gives the owner the right but not the obligation to modify the structure, either completely as in the case of demolition and redevelopment or partly as in the case of home improvements. Modeling property ownership in this manner has several advantages as it allows one to draw on standard implications of option theory. In a largely theoretical paper, Guthrie (2010), for example, argued that new house prices can exceed development costs even in the absence of restrictive zoning that limits development. They emphasized that in competitive land markets, uncertainty encourages land investors to delay development of vacant parcels of land.⁴³ In Guthrie's model, competition erodes the value of the development call option but does not reduce option value to zero (analogous to Bulan et al., 2009). Guthrie further argued that when housing demand is very price-elastic and interest rates are low, growing cities with heterogeneous quality vacant land should exhibit relatively high ratios of home price to development costs (including land and construction cost). Guthrie offered this as an alternative explanation for high price-to-cost ratios that have previously been attributed to restrictive zoning (e.g., Glaeser et al., 2005a,b) and topographic features that restrict land supply and development opportunities (e.g., Saiz, 2010).

Clapp and Salavei (2010) argued that as a structure ages and deteriorates, the value of the option to modify the structure increases. Clapp and Salavei (2010) emphasized that

⁴² Dye and McMillen (2007) used demolition permits to classify homes as those sold as teardowns and those for which the existing structure was retained. They also carefully modeled and tested for possible misclassification. Results indicate that misclassification occurs but only for a small share of homes and not enough to affect their results. These findings suggest that demolition permits are an effective way of classifying homes sold for demolition.

⁴³ Capozza and Helsley (1990) also discussed the manner in which uncertainty in the face of irreversible development increases the option value on undeveloped land and delays development. It is worth noting, however, that under alternate conditions, uncertainty can actually accelerate development, as in Bar-Ilan and Strange (1996) who considered the impact of lags between the time when an investor begins a project and that when it is completed.

increasing option value associated with older properties should increase property value, *ceteris paribus*. [McMillen and O'Sullivan \(2013\)](#) emphasized that as the time to redevelopment approaches, the value of existing structure attributes (e.g., bedrooms and size) should decline. At the point of redevelopment, as emphasized by [Rosenthal and Helsley \(1994\)](#) and [Dye and McMillen \(2007\)](#), existing structural attributes should have zero value.

[Clapp and Salavei \(2010\)](#) tested the implications of their model using home sales between 1995 and 2007 in Greenwich, Connecticut. They constructed alternative measures of the potential obsolescence of the existing structure, which should be positively related to the value of the option to modify or redevelop the home. Their preferred measure is the town's assessed value of the structure divided by the assessed value of the lot as if cleared of the structure. The intuition behind this measure is that structure and land are complementary goods. Families who are willing to pay a large amount for their location and other lot-specific attributes (e.g., square footage) will typically also want to occupy a valuable structure on the site. A low structure-to-lot value ratio, therefore, is characteristic of a home with a potentially dilapidated or obsolete structure that is ripe for renovation or redevelopment. Hedonic regressions of sale prices confirm that, controlling for many other attributes, low structure-to-value ratios increase sale price.

[Clapp and Salavei \(2010\)](#) also experimented with an alternative measure of structure obsolescence based on floor space of the home in question divided by floor space of recently built nearby homes. The intuition behind this measure is that newly built structures provide interior space that meets current preferences and demand. Existing homes of markedly different sizes are likely to be obsolete. This was graphically apparent in Vancouver, BC, in the late 1980s and 1990s. Small but relatively young homes were frequently demolished and invariably replaced with much larger homes (e.g., [Rosenthal and Helsley, 1994](#)). Using this alternate measure, [Clapp and Salavei \(2010\)](#) obtained results that are largely similar to those from their preferred measure.

An implication of these results is that the effect of house age on property value is more complicated and nonlinear than typically allowed for in traditional specifications of hedonic models that do not control for the value of the option to modify the structure. [Clapp and Salavei \(2010\)](#) interacted their measures of structure obsolescence with age of the home and obtained evidence consistent with this view. [McMillen and O'Sullivan \(2013\)](#) extended these ideas by emphasizing that at the time of sale, there is uncertainty as to whether a home will soon be renovated or redeveloped. This differs from [Rosenthal and Helsley \(1994\)](#) and [Dye and McMillen \(2007\)](#) who assumed that a home's teardown status is known with certainty at the time of sale. [McMillen and O'Sullivan \(2013\)](#) allowed for their more general view by estimating a hazard model of time to demolition following a sale. They then estimated separate hedonic models of sale price for homes likely to be redeveloped within 2 years of sale and those that are less likely to be redeveloped. Results confirm that uncertainty delays the decision

to redevelop as would be expected from standard option theory. In addition, structural attributes including especially interior space have much less effect on transactions price for homes destined for demolition and redevelopment within 2 years.⁴⁴

16.2.4.3 Filtering and gentrification

Home maintenance and the decision to redevelop a property clearly affect the pace and timing of change in neighborhood economic status. Here, we summarize results from two recent papers that offer evidence on the rate at which change occurs while shedding light on additional drivers of change.

Rosenthal (2008a) measured the change in neighborhood economic status on a decade-by-decade basis. As discussed earlier, economic status is measured as the average income in a census tract divided by average income across tracts in the MSA. For the balanced panel of census tracts drawn from the 35 cities previously described, the absolute value of the change in tract relative income ranges from a high of 14.9% from 1950 to 1960 to a low of 11.5% from 1970 to 1980. Similar values are obtained for a balanced panel of census tracts drawn from 331 metropolitan areas and followed from 1970 to 2000. For this sample, the absolute value of the change in tract relative income averages 12.2% from 1970 to 1980, 13.1% from 1980 to 1990, and 12.4% from 1990 to 2000.

Rosenthal (2014) takes a more microview by considering the rate at which individual homes filter down and up the economic ladder. Drawing on the AHS panel that follows homes—not people—over time, Rosenthal observes when homes turn over. The standard repeat sales model is then adapted to consider the change in real income of newly arriving occupants across turnover dates. More specifically, consider two successive turnovers when a home is age t and $t + \tau$ years, respectively. For each of these turnovers, income of the arriving occupant is written as

$$Y_t = e^{\gamma_t} f(\mathbf{X}_t; \boldsymbol{\beta}_t), \quad (16.6a)$$

$$Y_{t+\tau} = e^{\gamma_{t+\tau}} f(\mathbf{X}_{t+\tau}; \boldsymbol{\beta}_{t+\tau}), \quad (16.6b)$$

where $f(\mathbf{X}; \boldsymbol{\beta})$ is an unknown and potentially nonlinear function of the structural and neighborhood characteristics of the home ($\mathbf{X}$) and their shadow prices ($\boldsymbol{\beta}$). If $\mathbf{X}$ and $\boldsymbol{\beta}$ are time-invariant, taking logs and rearranging gives the log change in arriving occupant income between turnovers:

$$\log \left(\frac{Y_{t+\tau}}{Y_t} \right) = \gamma_{t+\tau} - \gamma_t + \omega_{t+\tau}, \quad (16.7)$$

where ω is a disturbance term and $f(\mathbf{X}; \boldsymbol{\beta})$ differences away. For a sample of properties ($i = 1, \dots, n$) that experience turnovers at various ages, an estimable version of (16.7) is

⁴⁴ See also Brooks and Lutz (2012) who argued that challenges with land assembly can create frictions that delay urban redevelopment.

$$\log\left(\frac{Y_{t+\tau,i}}{Y_{t,i}}\right) = \sum_{i=1}^{\tau_i} \gamma_t D_{t,i} + \omega_{t,i}, \text{ for home } i = 1, \dots, n, \quad (16.8)$$

where D_t equals -1 , 0 , or 1 depending on whether a given property of age t turns over for the first time, does not turn over, or turns over for the second time, respectively.

Rosenthal (2014) estimates expression (16.8) separately for rental and owner-occupied units. Results confirm that housing filters down with house age, on average, and more rapidly when the home is young. For rental units, the real income of an arriving occupant at a 50-year-old home is roughly 70% below that of the arriving occupant of a newly built unit, *ceteris paribus*. For owner-occupied housing, the difference is smaller, just 30%. On the whole, these estimates are large enough to account for the stylized fact that except for newly built subsidized units, lower-income families typically live in older homes.

To help explain the microfoundations of these rates of change, Rosenthal (2014) next considers the following simple housing demand function:

$$\log(h_{t,i}) = \theta_Y \log(Y_{t,i}) + \theta_q \log(q_{t,i}), \quad (16.9)$$

where housing (h) is treated as a package of homogenous quality adjusted units, rent per unit of housing on a quality adjusted basis is given by q , and the parameters θ_Y and θ_q are the income and price elasticities of demand for housing, respectively. Rearranging (16.9), differencing across turnover dates, and imposing a constant annual rate of depreciation d (i.e., $\log(h_{t+\tau,i}/h_{t,i}) = d\tau_i$) yield an alternate expression for $\log(Y_{t+\tau,i}/Y_{t,i})$:

$$\log\left(\frac{Y_{t+\tau,i}}{Y_{t,i}}\right) = \frac{d}{\theta_Y} \tau_i - \frac{\theta_q}{\theta_Y} \log\left(\frac{q_{t+\tau,i}}{q_{t,i}}\right) + \omega_{t,i}. \quad (16.10)$$

Expression (16.10) makes clear that filtering rates depend on the rate at which housing depreciates (d), the income and price elasticities of demand for housing (θ_Y and θ_q), and the rate of house price inflation ($\log(q_{t+\tau}/q_t)$). As would be expected, house price inflation slows the rate at which homes filter since $-\theta_q/\theta_Y > 0$ given downward-sloping demand ($\theta_q < 0$) and that the income elasticity of demand is positive ($\theta_Y > 0$). Because house price inflation differs across locations, filtering rates should as well. Moreover, if real house prices do not change as has been the long-run norm for most areas of the United States (see table 6 of Rosenthal, 2014), homes filter at a rate given by d/θ_Y . As noted earlier, Glaeser et al. (2008), Rosenthal (2014), and other studies confirm that θ_Y is well below 1. From expression (16.10), this amplifies the rate at which homes filter down for any given rate at which housing depreciates.

Rosenthal (2014) estimates (16.10) for rental and owner-occupied homes separately. House price inflation is measured using the Federal Housing Finance Association (www.fhfa.gov) repeat sale price index either as a direct proxy for the change in q or as an instrument treating the change in real rent (for rental units) and price (for

owner-occupied units) as endogenous controls. While estimates differ modestly with alternative specifications (e.g., OLS, 2SLS, and house fixed effects), results mostly indicate that rental units filter down at a real annual rate of roughly 2.5%, while owner-occupied filter down at a real annual rate of roughly 0.5%.⁴⁵

In summary, evidence suggests that neighborhoods change in economic status at a rate of roughly 12% per decade. Over a 50-year period, newly built rental homes experience a roughly 70% decline in occupant real income, while newly built owner-occupied units experience a 30% decline. Although mobile households might not recognize that their communities are in slow transition to a different status, the rates of change described above confirm that neighborhood economic status is dynamic and goes through long-running periods of rise and fall for most urban communities.

16.2.5 Does gentrification hurt local residents?

Gentrification: The buying and renovation of houses in deteriorated urban neighborhoods, especially by middle-income professionals, thus improving property values but often displacing low-income families.

The Random House College Dictionary, Revised Edition, Random House (1981)

Gentrification is controversial and often associated with negative connotations. The definition above explains why. Gentrification implies an asymmetry as higher-income families choose to move into a community, while lower-income households are at risk of being forced out because of rising property values and rents. This section reviews a small set of recent papers that consider who wins and who loses from gentrification and whether lower-income families are necessarily hurt.

Freeman (2005) used individual-level data from the Panel Study of Income Dynamics (PSID) to compare mobility rates of individuals residing in gentrifying communities to individuals in communities of similar economic status that do not gentrify. Freeman reported that mobility rates are similar for the two groups and from this concluded that gentrification does not disproportionately prompt out-movement of low-income families. McKinnish et al. (2010) conducted a complementary analysis using confidential long form data from the 1990 and 2000 US Census. They evaluated who moves out of gentrifying neighborhoods, who among existing residents remains in the community, and who moves in from other locations. Based on a detailed assessment of individual- and census tract-level data, McKinnish, Walsh, and White concluded that predominantly black low-income communities that gentrify attract middle-class black households.

⁴⁵ Expression (16.10) also allows one to estimate model-based measures of filtering rates by dividing separately estimated rates of depreciation (d) by separately estimated values for the income elasticity of demand for housing. Rosenthal (2014) performs this exercise and obtains additional, qualitative support for the idea that filtering rates increase with the rate at which homes depreciate and more so as the income elasticity of demand falls below 1.

In contrast, gentrifying communities with limited numbers of black families exhibit high rates of out-migration among black high school graduates.

Vigdor (2010) is the only study we are aware of that applies core economic principles to evaluate whether existing residents of gentrifying communities are hurt or benefit from the rising status of their neighborhood. Central to Vigdor's work is a sorting model in which households differ in their taste for neighborhood quality where neighborhood quality is exogenously given. In Vigdor's model, households with the weakest taste for neighborhood quality sort into the lowest-quality community in equilibrium. Households with the next strongest taste for neighborhood quality occupy second-tier-quality neighborhoods and so on up the quality ladder. In this setting, Vigdor showed that an exogenous decrease in quality in a given neighborhood lowers utility levels for all households in communities of equal or higher quality. This follows because of competition for access to given communities and the equilibrium adjustment of house prices.

Vigdor (2010) explored empirical implications of his model using a special feature of the AHS. Following homes over time, he carefully tracked the change in a set of indicators of neighborhood distress based on features within 300 feet of the housing unit. These include the number of abandoned buildings, bars on windows, streets in disrepair, and trash in the streets. Also included in the AHS is a qualitative variable in which respondents rate their satisfaction with their neighborhood on a 1–10 scale where 10 is the best. The AHS also allows one to observe changes in housing rents for rental units at the individual unit level.

Controlling for a variety of other factors, results indicate that improving neighborhoods signaled by a reduction in abandoned buildings and the like exhibit rising rents as would be expected. The question then is whether the willingness to pay for existing neighborhood improvements exceeds the increase in market rents. Such evidence would suggest that gentrification enhances welfare for current residents, while the opposite would be true if rent increases exceed willingness to pay. Based on a series of analyses and robustness checks, Vigdor reported that most renters are willing to pay between 1% and 3% of their annual income for a one standard deviation increase in neighborhood quality, while the change in rent from improved neighborhood quality is much less. From this assessment, Vigdor concludes that, at least for many renters, improved neighborhood quality arising from gentrification enhances welfare.

From a qualitative perspective, Vigdor's analysis reminds us that rent increases are a poor proxy for the welfare effects of gentrification on existing residents who may themselves benefit from improved neighborhood amenities. Indeed, the possibility that gentrification may be good for local residents has been raised in various popular press reports, an example of which is a recent article in *New York Magazine* (see Davidson, 2014, "Is Gentrification All Bad?"). Nevertheless, controversy remains as is evident in a *New York Times* article in 2008 by Williams (2008), "Harlem Area Is Blighted, State Agency Declares."

16.2.6 US context versus Europe

In [Section 16.1](#), we noted that most of the literature on neighborhood dynamics has been written in a US context, and for that reason, our review reflects this. Here, we pause to briefly consider the possible influence of two prominent contextual features of US cities on neighborhood income dynamics and in comparison with cities found in Europe.

US cities are young compared to most European cities. Partly for that reason, few US cities have the sort of historic downtowns that are characteristic of cities like Madrid, Barcelona, Paris, Basel, Rome, or Vienna, not to mention truly ancient cities like Jerusalem in the Middle East. In many of these places, a protected historic center provides a unique urban amenity that may attract higher-income families and has the potential to slow the tendency for filtering and related cycles in economic status. The paper by [Brueckner et al. \(1999\)](#) on Paris provides evidence of this. In that sense, historic city centers may have much the same dampening effect on neighborhood change as attractive physical amenities that have recently been studied by [Lee and Lin \(2013\)](#). Partly for that reason, neighborhood income dynamics may be more pronounced in the United States relative to Europe, but this remains an open question.

A second feature of the United States that may have also amplified neighborhood income stratification and related dynamics is its history of racial tension and the Great Migration. As described in [Cutler et al. \(1999\)](#), between 1890 and 1970, the United States experienced a massive flow of predominantly poor African-Americans from the rural south to cities throughout the country, especially in the south and in the industrial northeast and Midwest. That migration gave African-Americans access to higher-paying manufacturing jobs and contributed to prosperous African-American neighborhoods, such as Harlem as described in [Section 16.1](#). However, post World War II, the Great Migration was accompanied by equally dramatic suburbanization of both white households ([Baum-Snow, 2007](#); [Boustan, 2010](#)) and manufacturing ([Glaeser and Kahn, 2001](#)). By the 1970s, US manufacturing was in sharp decline and migration from the rural south had slowed to a trickle leaving highly segregated central cities with declining economic fortunes and population. It is certain that the massive migration of lower-income, southern rural blacks into US cities increased racial segregation in US cities. It seems likely that this also contributed to income segregation and amplified cycling of neighborhood economic status over the long sweeps of time described earlier. However, this is also an open question.

16.3. CITY DYNAMICS

16.3.1 Overview

This section reviews a particular slice of the literature on city dynamics. As with the previous section, our focus is on the extent and conditions under which a location's

economic status relative to other locations is entrenched or changes over time in a systematic or nonsystematic fashion. An important conclusion from this section is that when metropolitan area status is measured based on population size, as has been the case in most of the agglomeration literature, MSA status appears to be extremely stable. However, when status is measured based on direct indicators of economic activity and especially *per capita* income, MSA economic status is far more dynamic than has generally been recognized.

As with the previous section on neighborhoods, we begin with four stylized facts, the first three of which are taken from the literature, while the fourth is new to this review. We then discuss various recent papers that help to explain and interpret the stylized facts while drawing out implications for MSA-level dynamics.

16.3.2 Three previous stylized facts and one new observation

16.3.2.1 *Stable city size distributions that approximate the rank-size rule*

Eaton and Eckstein (1997) examined changes in relative city sizes for 40 long-established cities in Japan between 1876 and 1990 and 39 cities in France between 1925 and 1985. They found that most cities exhibit parallel population growth. Moreover, the city size distributions are consistent with the rank-size rule and Lorenz curves confirm that the size distributions of French and Japanese cities remain stable over time. Sharma (2003) examined the size distribution of cities in India over a roughly 100-year period. She reported short-run deviations from long-run parallel growth paths but those deviations tend to fade after less than a decade. Her results also confirm that city size distributions tend to be stable.

Probably, the most dramatic evidence of stable spatial distributions of population comes from Davis and Weinstein (2002). Drawing on current, historical, and archaeological data for Japan, they evaluated the extent to which the spatial distribution of population across 39 regions throughout Japan adheres to the rank-size rule at different points in time over an 8000-year history. For each period, they estimated

$$\text{Log}(\text{population rank}) = \text{Constant} - b \log(\text{population size}), \quad (16.11)$$

where a coefficient for b of -1 implies that the n th largest location is $1/n$ the size of the most populous area, in which case Zipf's law and the rank-size rule are said to hold.

Table 16.4 reproduces estimates of b from table 1 of Davis and Weinstein (2002, p. 1273) for 12 periods from -6000 BC to 1998. The patterns are stunning with Zipf coefficients close to -1 in nearly all periods. These and other estimates in the literature confirm the first stylized fact that the spatial distribution of population within a country tends to be extremely stable over long sweeps of time and approximate the rank-size rule. This feature of economic growth has been appreciated for some time now in the literature. See, for example, Gabaix and Ioannides (2004).

Table 16.4 8000 Years of Zipf's law in Japan^a

Period (year)	Zipf coeff.	Period (year)	Zipf coeff.
–6000 to 300	–0.809 (0.217)	1600	–1.192 (0.068)
–300 to 300	–1.028 (0.134)	1721	–1.582 (0.113)
725	–1.207 (0.133)	1798	–1.697 (0.120)
800	–1.184 (0.152)	1872	–1.877 (0.140)
900	–1.230 (0.166)	1920	–1.476 (0.043)
1150	–1.169 (0.141)	1998	–0.963 (0.025)

^aReproduced from Davis and Weinstein (2002, table 1, p. 1273), “Zipf coefficient” column. Standard errors in parentheses.

16.3.2.2 Slow movement of individual cities up and down the city size distribution

Returning to Eaton and Eckstein (1997), for both Japan (1876–1990) and France (1925–1985), they divided the cities in their samples into six relative size categories based on a fixed reference point. In both samples, Eaton and Eckstein (1997) found that there is relatively little tendency for cities to transition to a different size category over the sample horizon. In Japan, 0–22% of cities transition to a different size category depending on the initial category to which a city is assigned. In France, the corresponding range is 10–31%. Analogous results are obtained by Black and Henderson (2003) for 282 US metropolitan areas between 1900 and 1990 (and five city size categories). As highlighted in Section 16.1, Duranton (2007) documented that of the 10 largest metropolitan areas in the United States in 1977, the average change in rank in the MSA rank-size distribution over the following 20 years was just 1.2. Moreover, only 2 of these 10 MSAs, Cleveland and San Francisco, experienced a change in rank beyond one position in the MSA size distribution.

Broadly speaking, these and other studies in the literature provide evidence in support of a second stylized fact: cities tend to move only very slowly up and down the city population size distribution. Together, these first two stylized facts could suggest that the economic status of individual MSAs hardly changes even over extended periods of time. The next two stylized facts suggest that such an interpretation would not be correct.

16.3.2.3 Fast changes in city industrial mix relative to total employment

In an important departure from earlier literature on urban dynamics, Duranton (2007) focused on the pace at which the mix of local economic activity changes over time. He showed that metropolitan areas in both France and the United States exhibit comparatively rapid changes in their industrial mix over time in comparison with population

size. In a follow-up study, [Findeisen and Südekum \(2008\)](#) documented similar patterns for West Germany.

In both [Duranton \(2007\)](#) and [Findeisen and Südekum \(2008\)](#), industry churning across MSA's is measured by comparing the degree to which employment shares in individual industries within a given metropolitan area change over time. *Churn* is thus measured as

$$Churn_c = \frac{1}{25} \left(\sum_{t=1977}^{2001} \sum_{z=1}^n \frac{|e(z, c, t+1) - e(z, c, t)|}{e(c, t)} \right) \text{ for } c = 1, \dots, 326, \quad (16.12)$$

where e is the level of employment in industry z in city c , at time t . In [Findeisen and Südekum](#), *Churn* is measured using 25 years of data from 1977 to 2001 for 326 MSAs in West Germany. Similar magnitude samples were used by [Duranton \(2007\)](#) for the United States and France. In both studies, values for *Churn* were then compared to the yearly average aggregate change in each MSA's total employment and also the average annual change in each industry's share of national employment.

[Table 16.5](#) reproduces estimates of *Churn* from table 1 of [Findeisen and Südekum \(2008\)](#). The estimates include measures for the United States and France taken from [Duranton \(2007\)](#) and also for West Germany as calculated by [Findeisen and Südekum \(2008\)](#). Two patterns are especially striking. First, the average value of *Churn* across MSAs is remarkably similar for West Germany, the United States, and France. Second, the industrial mix within a given MSA changes at nearly twice the rate necessary to accommodate change in the MSA's aggregate-level employment. These patterns indicate that the industrial mix within individual MSAs tends to change rather quickly in comparison with changes in the MSA's total level of employment, our third stylized fact.

16.3.2.4 Very fast movement of cities up and down in economic status based on income

The comparatively fast change in an MSA's industrial mix is suggestive that MSA economic status may be subject to much more change over time than implied by stable MSA population size distributions as documented in the first two stylized facts. This section introduces a fourth stylized fact that reinforces that view.

Table 16.5 Industry movements across metropolitan areas

Period (year)	$Churn_c$	ΔEmp_c	$Churn_c / \Delta Emp_c$	$\Delta SecEmp$
West Germany ^a	4.98%	2.29%	2.17	2.62%
The United States ^a	8.26%	4.10%	2.01	≈ 5%
France ^a	11.40%	5.20%	2.19	≈ 5%

^aSource: Reproduced from [Findeisen and Südekum \(2008, table 1, p. 329\)](#). Values for West Germany were calculated by [Findeisen and Südekum](#). Values for the United States and France were taken from [Duranton \(2007\)](#).

In Figure 16.6, we display two Lorenz curves that summarize inequality across US metropolitan areas in 2000.⁴⁶ The solid line is based on differences in population across MSAs, treating total population across all metropolitan areas in the United States as the normalizing factor. The dashed line is calculated in an analogous manner and is new to the literature. This line reports inequality in median income across MSAs treating the sum of median income across all metropolitan areas as the normalizing factor (with each MSA given equal weight). The MSA population size distribution displays a familiar highly skewed pattern that clearly dwarfs the degree of inequality in median income across metropolitan areas.

It should come as no surprise that the distribution of median income across MSAs is much flatter than its population analogue. Glaeser and Mare (2001) reported that nominal average wage rates in 1990 were roughly 30% higher in MSAs with over 1 million people as compared to rural areas outside of MSAs. That is a small difference compared to the difference in population density between such areas. Moreover, the urban wage premium falls dramatically after controlling for worker attributes. In a review of earlier literature, Rosenthal and Strange (2004) reported that doubling city size increases wage by 3–5% controlling for worker attributes (consistent with estimates in Rosenthal and Strange, 2008). More recent literature reviewed by Combes and Gobillon (this volume) suggests an even smaller urban wage premium, possibly as low as 1–3%. This does not, however, mean that differences in income distributions across MSAs are fixed.

Using the public use microdata sample of the Decennial Censuses (see www.ipums.org), we examine changes in an MSA economic status between 1980 and 2010 using a balanced panel of 219 MSAs in the United States. For these purposes, an MSA economic status is measured based on its rank among the other metropolitan areas when comparing income at a given point in the within-MSA income distribution. Results are reported in Table 16.6 based on several different measures of MSA income. Reading across the columns, these include 10th percentile income, 25th percentile income, 50th percentile income, 75th percentile income, and 90th percentile income. For each column, the distribution of the 1980–2010 change in rank among the 219 MSAs is reported going down the rows, from the 10th percentile to the 90th percentile. By construction, the mean change in rank must always be zero as shown in the third row. The median change in rank (in the fourth row) is always close to zero and suggestive of an approximately symmetrical distribution. This accounts for the negative values in the first two rows (for the 10th and 25th percentiles) and the positive values in the last two rows (for the 75th and 90th percentiles). For reference, the final column in the table reports analogous measures based on 1980–2010 change in MSA population rank.

The patterns in Table 16.6 indicate much larger changes in MSA economic rank based on the different measures of income as compared to population. As an example,

⁴⁶ The Lorenz curves are based on the authors' calculations using public use microdata from the 2000 Decennial Censuses as obtained from www.ipums.org.

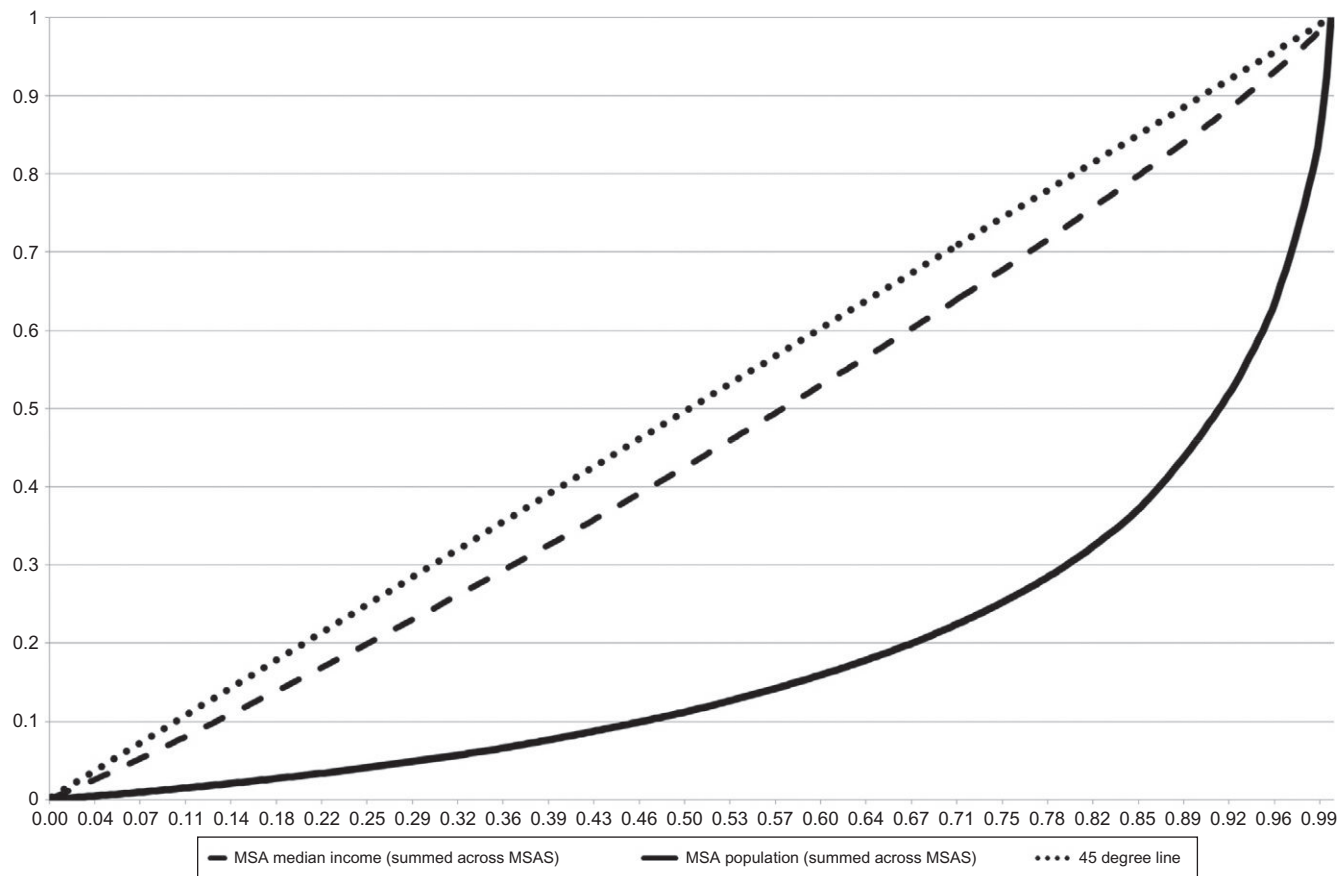


Figure 16.6 Lorenz curve measures of inequality across MSAs in median income and populations.

Table 16.6 Change in MSA income rank between 1980 and 2010^a

	MSA income measured at percentile					MSA population
	10th	25th	50th	75th	90th	
10th pctl	−67	−71	−69	−64	−64	−28
25th pctl	−36	−30	−32	−33	−29	−10
Mean	0	0	0	0	0	0
Median	−3	−1	−4	−5	−1	1
75th pctl	34	31	30	25	26	14
90th pctl	65	69	87	78	67	26

^aAuthor calculations based on individual-level data from the 1980 and 2000 US 5% public use microdata sample (PUMS) obtained from www.ipums.org. All calculations based on a balanced panel of 219 MSAs.

at the 25th percentile as indicated in the second row, MSA population rank falls 10 positions. In contrast, rank based on income falls roughly 30–35 positions depending on the income measure used. This suggests a new fourth stylized fact: based on various measures of *per capita* income, metropolitan areas move very fast up and down in economic status relative to other MSAs, especially in comparison with population rank.

16.3.3 Drivers of persistence in metropolitan economic status

In this section, we highlight three drivers of persistence in metropolitan-level economic status. These are natural advantages of the location, self-reinforcing effects of agglomeration and economic activity, and the enduring effects of culture. Each is discussed below.

16.3.3.1 Natural advantages

Some of the most striking evidence of extreme persistence is in a series of papers that examine the effects of allied bombings in Japan and Germany during World War II on the long-run population of cities in these two countries. As noted earlier, [Davis and Weinstein \(2002\)](#) examined the distribution of population across regions throughout Japan going back 8000 years. From the Stone Age period to AD 1200, the correlation between past and current (based on 1998) region-level population is roughly 0.5–0.6 with higher correlations after AD 1200. These patterns indicate that areas that were important thousands of years ago remain so today.

[Davis and Weinstein \(2002\)](#) also reported that the correlation between city-level populations in Japan between 1998 and 1920 is 0.94. In some respects, these patterns are even more striking than the 8000-year correlations since many of the Japanese cities were heavily bombed during World War II.⁴⁷ The postwar patterns indicate that despite

⁴⁷ In the 66 Japanese cities covered in the United States Strategic Bombing Survey, estimates indicated that over half of all structures were destroyed and that some cities lost as much as 50% of their population during the war.

widespread destruction during the war, that damage had little long-term impact on the spatial distribution of population in Japan as the country rebuilt and recovered. Two especially prominent examples of this sort of extreme tendency toward mean reversion are Hiroshima and Nagasaki, which were devastated by the nuclear detonations. [Davis and Weinstein \(2002\)](#) showed that although both cities suffered dramatic loss in population during the war, Nagasaki returned to its long-run population growth path by 1960, while Hiroshima returned to its long-run population growth path by 1975 (see [Davis and Weinstein, 2002](#), figure 2, p. 1282).

In a subsequent paper, [Davis and Weinstein \(2008\)](#) further showed that the industrial mix of economic activity within individual metropolitan areas in Japan also displays considerable mean reversion and long-run resilience to damage incurred during World War II. The loss of manufacturing and other industry-specific employment during World War II was followed by substantially faster growth in war-damaged industries and a strong tendency for individual cities to return to their former, prewar industrial composition. Based on these patterns, [Davis and Weinstein \(2008\)](#) concluded that location fundamentals (i.e., natural advantages) play an important role in determining the distribution of population and industrial mix over space.

[Brakman et al. \(2004\)](#) conducted a similar exercise for housing stock in German metropolitan areas prior to and following World War II. In West Germany, they found that cities that lost the most housing stock during Allied bombings displayed the fastest postwar growth in housing stock. Mirroring much of Japan, West German metropolitan areas returned to their previous long-run trends by the 1960s. In contrast, in East Germany, postwar spatial patterns of growth in housing stock were mostly unrelated to the level of wartime damage. [Brakman et al. \(2004\)](#) suggested that central planning and limited property rights in postwar East Germany may account for the East–West difference in postwar recovery patterns. However, they also cautioned that their sample of East German cities is relatively small and that idiosyncratic features of the split between East and West Germany may confound interpretation of postwar patterns of development.

Recent work by [Glocker and Sturm \(2013\)](#) sheds further light on these patterns by examining prewar German cities that became part of Poland after the war. Efforts by West Germany to rebuild its cities following the war cannot be the driving factor behind growth in these formally German cities. Similarly, a tendency for displaced population to return to their former and now destroyed cities in order to renew previous social networks would not seem a likely driver of growth in formally German metropolitan areas. The same is true for any sense of patriotic duty that might foster a desire to rebuild an area that was destroyed. Indeed, the change in border for formerly German cities incorporated into postwar Poland invalidated previous property rights and prevented the return of displaced prewar population. Nevertheless, [Glocker and Sturm \(2013\)](#) also found that bombed cities fully recover their prewar population shares. On balance, therefore, the evidence for Germany is consistent with [Davis and Weinstein's \(2002, 2008\)](#) findings

for Japan. In both countries, cities regained their prewar relative status in population rank and in a manner that suggests an important role of location fundamental advantages in shaping the spatial distribution of economic activity.⁴⁸

16.3.3.2 Agglomeration economies

A completely different argument for extreme persistence of a city's economic status is based on self-reinforcing spillovers from agglomeration. Imagine a featureless plain and that economic activity is dropped first at site A instead of site B. If agglomeration fosters productivity spillovers, then site A gains an immediate productivity advantage over site B that is self-reinforcing unless or until congestion costs slow site A's growth (as in [Cuberes, 2011](#), for example). Under these conditions, site A will forever be a larger and likely higher *per capita* income city than site B because of greater productivity even though the events that initiated development at site A may no longer be relevant.

[Bleakley and Lin \(2012\)](#) made creative use of data on historical patterns of development that allows them to look for evidence of the sort of persistence just described. They examined development patterns adjacent to canoe portage paths that once provided economically important routes around rapids and waterfalls along natural fall lines that occur at the edge of the Appalachian escarpment in the Eastern United States. These portage paths were once an important trade route that contributed to local concentrations of population and equipment in order to facilitate overland transportation of goods. As such, portage paths prompted the development of various small towns along the Appalachian Mountain chain. However, portage paths lost their economic relevance with the advent of rail transportation and have no intrinsic commercial value today except for recreation.

In the absence of self-reinforcing agglomeration economies, we should not see unusual concentrations of population and economic activity adjacent to portage paths today. [Bleakley and Lin \(2012\)](#) found the opposite and documented that portage path locations have significantly higher employment density today. They examined several potential mechanisms for this result including the influence of railroads, education, industrial composition, and the early use of water power. Only the early use of water power survives the inclusion of controls for other historical factors. Portage cities appear to have adopted water power earlier than other locations along the same rivershed and so obtained an advantage that leads to future population growth. More generally, the patterns in [Bleakley and Lin \(2012\)](#) support the view that agglomeration economies can contribute to extreme persistence in a city's economic status.

In an analogous set of papers, [Jedwab and Moradi \(2014\)](#) and [Jedwab et al. \(2014\)](#) examine the impact of colonial railroads in Africa as a whole, as well as in Ghana and Kenya. Both papers show that access to colonial railroads implies higher postcolonial

⁴⁸ Consistent with the above literature, [Acemoglu et al. \(2005\)](#) argued that the rise of Western Europe after 1500 was disproportionately driven by growth among countries adjacent to the Atlantic Ocean.

population levels even though colonial railroads quickly fell into disrepair and disuse throughout Africa following independence from colonial rule. In Africa as a whole and in Ghana, there is no evidence that this impact operates through roads that were built along original railroad paths. Instead, population patterns in 2000 are entirely explained by urbanization in 1960 when colonial rule was still the norm.

The connection between year 2000 and colonial era population patterns is robust to a variety of identification strategies including the use of straight line distance from major ports to historical capitals as an instrument for railroads. In Kenya, the authors show that access to colonial railroads increases present-day population, urbanization, and access to public services and decreases poverty. The authors argue that the railroads were built for geopolitical purposes on the continent and had little to do with local economic factors within Kenya. Further, no population effects are found for placebo lines based on branch lines that were planned, but not built. Like [Bleakley and Lin \(2012\)](#), these papers suggest that obsolete transportation networks have an enduring effect on current spatial patterns of economic activity, consistent with the view that localized shocks to growth have long-term effects.⁴⁹

16.3.3.3 Culture

A different mechanism that can contribute to extreme persistence in economic development at the regional level is culture. [Tabellini \(2010\)](#) showed that economic growth between 1995 and 2000 within European subregions is related to cultural factors after controlling for country fixed effects, school enrollment in the 1960s, and urbanization in 1850 as past indicators of economic development. The influence of culture also persists when Tabellini used historical variables to instrument for culture, including literacy in 1880 and political institutions between 1600 and 1850. In another recent example, [Falck et al. \(2012\)](#) examined linguistic data from 1879 to 1888 in German schools. They showed that region-level migration decisions are heavily influenced by linguistic similarity even after controlling for geographic features, travel distance, religious boundaries, historical industrial structure, and jurisdiction borders. As with portage paths in the United States and colonial railroads in Africa, literacy rates, political institutions, and linguistic differences from over 100 years ago are capable of fostering extreme persistence in spatial patterns of economic activity.

⁴⁹ Also consistent with a self-reinforcing role for agglomeration economies, [Hanlon and Miscio \(2013\)](#) found that the interruption of the cotton trade caused by the US Civil War had a significant negative impact on employment in English towns tied to the textile industry for many decades after the end of the war. [Wahl \(2013\)](#) found that GDP is higher at locations at or close to the site of a medieval-period trade center even after instrumenting for trade center status with controls for geography, climate, and the historical presence of Christian populations.

16.3.4 Drivers of change in metropolitan economic status

In this section, we discuss drivers of change in metropolitan-level economic status. These include productivity shocks that affect labor demand as well as amenities that affect labor supply. We also highlight the role of housing supply that, when inelastic, has potential to contribute to and in some instance amplify metropolitan area *per capita* income dynamics.

16.3.4.1 Productivity and labor demand

There has been a strong and well-placed tendency in the literature to equate city growth in population and employment with rising productivity and economic status. Much of that literature has been reviewed elsewhere (e.g., [Duranton and Puga, 2004](#); [Rosenthal and Strange, 2004](#); [Behrens and Robert-Nicoud, this volume](#); [Combes and Gobillon, this volume](#); [Carlino and Kerr, this volume](#)). In this section, we highlight a particular slice of recent additions to that literature that have direct implications for the stylized facts outlined above.

An old literature, beginning with [Losch \(1954\)](#) and [Christaller \(1966\)](#), often referred to as *Central Place Theory*, outlines a largely static model of a hierarchy of cities within a system of cities framework. In its simplest form, nonexport industries differ in their internal economies of scale and also with respect to *per capita* demand for their services.⁵⁰ Allowing for travel costs, industries with deep internal economies of scale and low *per capita* demand typically thrive only in large metropolitan areas with large numbers of nearby customers. An example would be specialized medical facilities. Industries with low internal economies of scale and high *per capita* demand thrive in metropolitan areas of all sizes, as with gas stations.

An implication of the central place theory model is that large metropolitan areas are not simply scaled-up versions of small cities but instead have more diverse economies. That diversity has dynamic implications in that it can contribute to cross-fertilization of ideas, innovation, and growth as emphasized by [Jacobs \(1969, 1984\)](#) and explored in various papers in the literature on agglomeration economies (see [Rosenthal and Strange, 2004](#), for a review). [Duranton and Puga \(2001\)](#), for example, developed an explicitly dynamic model in which large diverse cities serve as incubators and nurture newly created industries. An implication is that industries benefit disproportionately from large-city locations when they are young and are more likely to relocate to more specialized locations as the industry matures. Consistent with this view, [Duranton and Puga \(2001\)](#) reported that 72% of French plant relocations entail moves from diverse to more specialized locations. In especially innovative sectors of the economy, these sorts of relocations are even more extreme: 93% for research and development and 82% for pharmacology.

⁵⁰ See [O'Sullivan \(2003\)](#), chapter 5, pp. 92–118, for a discussion of this model.

Duranton (2007) developed a conceptual model that builds off of this earlier work and that generates predictions that fit the first three stylized facts described above. He modified Grossman and Helpman's (1991) model of innovation and growth by adding cross-industry spillovers in a manner that allows for innovations within and across cities. In Grossman and Helpman (1991), research firms compete within each industry to develop new technology for a given industry. Whichever firm succeeds becomes the next technology leader in the industry and only that firm can profitably operate in the industry. In Duranton (2007), a firm in one industry may generate a discovery that makes it the technology leader in a different industry. Further, research firms in an industry can only operate in the city in which the technology leader for that industry is located. In this way, cross-industry spillovers allow cities to both gain and lose industries over time.

Duranton showed that his model leads to comparatively rapid churning of the industrial composition within individual cities, consistent with the third stylized fact above. He also showed that with multiple industries present in each city, individual cities tend to experience slow economic decline or growth as the number of industries present in the city changes over time, consistent with the second stylized fact above. Finally, using parameter values calibrated to fit the model to data for the United States and France, simulations largely replicate the existing size distribution of metropolitan areas in those countries and approximate Zipf's law. Analogous results are obtained by Findeisen and Südekum (2008) for Germany.⁵¹

An implication of Duranton's (2007) model that is not fully developed in his paper or in Findeisen and Südekum (2008) is that the evolution of individual city size should be a mean reverting process.⁵² A simple derivation of the expected change in city size conditional on a city's initial period size confirms this result. As in Duranton (2007), there are m cities in the system, each of which has one first-nature industry that cannot move. In addition, there are n mobile (not first-nature) industries spread throughout the system of cities. All industries are of equal size (employment) so that a city's size is completely determined by the number of mobile industries present plus its one first-nature industry. All individual industries—both first-nature and mobile industries—generate own-industry innovations with probability D and cross-industry innovations in any industry with probability C , with $D > C$. Importantly, in a given period, it is never possible for a given industry to receive more than one innovation. Instead, a given industry receives either zero or one innovation.

⁵¹ See Gabaix and Ioannides (2004) and Desmet and Henderson (this volume) for in-depth discussions of Zipf's law and the size distribution of cities. Outside of the economics literature, there is also recognition that history-dependent stochastic processes of the sort modeled by Duranton (2007) yield Zipf's law rank distributions of outcomes for many natural and social systems (see Corominas-Murtra et al., 2014, for example).

⁵² For macroeconometric evidence on convergence across states and regions in the United States, see Holmes et al. (2013), Mello (2011), and Choi (2004).

Consider now a city p in period t with η_{pt} mobile industries. To determine the expected size of p in $t+1$ conditional on its period- t size, we first calculate the probability that a given mobile (not first-nature) industry q will be present in the city in $t+1$. Summing probabilities over all mobile industries $q=1, \dots, n$, gives

$$E\left[\eta_{p(t+1)} \mid \eta_{pt}\right] = \sum_{q=1}^n \text{Prob}\left(\text{industry } q \text{ in city } p \text{ in } t+1 \mid \eta_{pt}\right). \quad (16.13a)$$

Next, divide this summation into two parts as shown in expression (16.13b). The first summation reflects the expected contribution of city p 's initial set of η_{pt} mobile industries to the number of mobile industries present in p in the subsequent period. For a given industry q , this equals the probability that no industry anywhere in the system of cities develops an innovation in q between periods *plus* the probability that an innovation in q does occur but is generated by one of the industries already present in p . Summing over all η_{pt} mobile industries in p then gives the expected number of mobile industries in t that are still present in $t+1$. The second summation in (16.13b) reflects the expected contribution of the $n - \eta_{pt}$ mobile industries not in p at time t to the number of industries present in p at time $t+1$:

$$\begin{aligned} E\left(\eta_{pt+1} \mid \eta_{pt}\right) &= \sum_{q=1}^{\eta_{pt}} (\text{Prob}(\text{no innovation in } q) + \text{Prob}(\text{innovation in } q \text{ generated in } p)) \\ &\quad + \sum_{q=\eta_{pt}+1}^n \text{Prob}(\text{innovation in } q \text{ generated in } p \mid q \text{ not in } p \text{ at } t). \end{aligned} \quad (16.13b)$$

Using the notation above, the probability of no innovation in one of p 's initial mobile industries (the first term in the first summation in (16.13b)) can be written as $1 - D - (n + m - 1)C$, where $n + m - 1$ are the number of industries in the system apart from the own-industry including both mobile and first-nature industries. The probability of an innovation in a specific one of p 's initial mobile industries from within p is given by $D + \eta_{pt}C$ (the second term in the first summation in (16.13b)). The probability of an innovation in a mobile industry not initially in p but which is generated from within p is given by $(\eta_{pt} + 1)C$ (the second summation in (16.13b)). Substituting these expressions into (16.13b), the expected number of mobile industries in city p in $t+1$ can be written as

$$\begin{aligned} E\left(\eta_{pt+1} \mid \eta_{pt}\right) &= \left[\sum_{q=1}^{\eta_{pt}} (1 - D - (n + m - 1)C) + (D + \eta_{pt}C) \right] \\ &\quad + \left[\sum_{j=\eta_{pt}+1}^n (\eta_{pt} + 1)C \right]. \end{aligned} \quad (16.13c)$$

Subtracting η_{pt} from (16.13c) gives the expected change in city size between periods, which, upon collecting terms, simplifies to⁵³

$$E(\eta_{pt+1}|\eta_{pt}) - \eta_{pt} = (-\eta_{pt}m + n)C. \quad (16.13d)$$

Expression (16.13d) indicates that there is a unique steady state $\eta^* = n/m$ at which the expected change in city size between periods is zero. This steady state is “stable” because when city size is below that level, expression (16.13d) indicates that the city is expected to grow, while the reverse is true when city size is above the steady state. It should be noted, however, that this result requires that C be a constant. If instead C was sensitive to the number of industries in the city, possibly because of external scale economies in research and innovation, then C would be a nonlinear function of city size. In this instance, multiple steady states could exist.

Interestingly, [Duranton and Puga \(2001\)](#) postulated such a model in their Nursery City paper where young, innovating industries benefit from being in large cities with a diverse mix of industries. Further, if C varies with the type of innovating industry, then C varies with the mix of industries present in the city and depends on the city’s economic and industrial history. This likely increases the potential for multiple steady states and reduces the potential for mean reversion.

For these reasons, the [Duranton \(2007\)](#) model has very different implications for long-run population growth trends across cities as compared with the model in [Eaton and Eckstein \(1997\)](#), which predicts parallel growth across cities. Empirically, [Findeisen and Südekum \(2008\)](#) did find some evidence of mean reversion in total employment among growing German metropolitan areas, a pattern that is inconsistent with parallel growth. Nevertheless, [Eaton and Eckstein \(1997\)](#), [Black and Henderson \(2003\)](#), and [Sharma \(2003\)](#) all reported patterns roughly consistent with long-run parallel growth for metropolitan areas in Japan, the United States, and India, respectively, as did [Desmet and Rappaport \(2013\)](#) for US counties after the 1940s.⁵⁴

It is also interesting to note that industrial churning and change in MSA-level *per capita* income are both rapid in comparison with MSA-level change in population. Resolution of this difference requires going beyond [Duranton’s \(2007\)](#) conceptual model, which

⁵³ Expression (16.13c) simplifies to $(\eta_{pt+1}|\eta_{pt}) = \eta_{pt} + \eta_{pt}(-(n+m-1) + \eta_{pt})C + (n - \eta_{pt})(\eta_{pt} + 1)C$.

⁵⁴ [Cuberes \(2011\)](#) developed a model of sequential city growth that also ultimately implies parallel growth. In the Cuberes model, capital initially flows to the largest city, which offers greater returns because of productivity spillovers associated with agglomeration. As excess returns in the largest city diminish, capital flows to the next largest city, and so on down the line toward smaller, newer locations. In any given period, growth is concentrated among a small number of cities that grow in size relative to other cities in the system. Using data from numerous countries, Cuberes provided empirical support for his model. An implication of Cuberes framework is that with sufficient passage of time, all cities within a country will have achieved their steady-state relative population size and should exhibit parallel growth rates thereafter.

implies that MSA-level *per capita* income is constant across metro areas and over time given the assumption that all industries are alike. For *per capita* income to vary with city size, there must be a correlation between the quality and quantity of industries within a given city. Such correlations are implied by [Eaton and Eckstein \(1997\)](#) and much of the literature on agglomeration economies, which tends to argue that larger cities are more productive, an implication of which is that *per capita* income and city size should move together. Returning to the patterns in [Table 16.6](#) (our fourth stylized fact), evidence that MSA *per capita* income changes more rapidly than population and seemingly more rapidly than industrial mix is suggestive that something is going on in metropolitan areas beyond what has been recognized in the literature.

16.3.4.2 Amenities and labor supply

This section considers a very different driver of change in MSA-level *per capita* income and related economic status. Local amenities and the perceived quality of life have a direct effect on household utility and the desire of families to live and work in a given metropolitan area. This affects MSA-level labor supply, equilibrium wages, and the economic status of the metropolitan area. A substantial literature on urban quality of life has emphasized these points, including [Roback \(1982\)](#), [Blomquist et al. \(1988\)](#), [Gyourko and Tracy \(1991\)](#), [Gyourko et al. \(1999\)](#), [Chen and Rosenthal \(2008\)](#), and [Albouy \(2008, 2009\)](#).⁵⁵ A central tenant of the quality of life literature is that *real wages* adjust across metropolitan areas to compensate mobile workers for differences in MSA-level amenities. We focus here on recent additions to this literature that have especially direct implications for changes in MSA-level economic status as reflected in indicators of *nominal per capita* income.

The evidence that households are drawn to high-amenity cities is compelling. [Glaeser et al. \(2001\)](#), for example, found that high-amenity cities have grown faster in population than low-amenity cities and that urban rent has risen faster than urban wages, suggesting a growing premium for residing in attractive locations. [Chen and Rosenthal \(2008\)](#) documented that retirees are especially likely to migrate toward high-amenity locations that are not necessarily attractive to business (as with Tampa Bay but not San Francisco) as such locations provide high quality of life at lower costs. Nevertheless, these patterns by themselves do not ensure that high-amenity cities and especially larger cities exhibit high levels of *per capita* income.

[Glaeser et al. \(2001\)](#) and [Quigley \(1998\)](#) both argued that city size is extremely important for facilitating consumption, consistent with the diversity of activity implied by central place theory as described at the outset of the previous section. High-end restaurants, theater, and various other cultural amenities typically require large numbers of patrons to bring down average cost and for that reason thrive best in larger metropolitan areas

⁵⁵ [Gyourko et al. \(1999\)](#) reviewed much of the early literature in this area.

with their larger populations. To the extent that such cultural amenities appeal to higher-income families, this draws higher-skill, higher-income workers to larger cities. Various studies, however, also make clear that the link between city size and the quality of consumption opportunities is not restricted to just higher-income households.

As discussed earlier, [Waldfoegel \(2008\)](#) showed that diversity of restaurant options increases with the size of the local population including the presence of both high- and lower-cost venues. Similarly, [Waldfoegel \(2003\)](#) demonstrated that the diversity of radio broadcast programs increases with the number and size of the different target audiences. Because minority communities tend to be larger in larger cities, such locations also offer a greater variety of radio broadcast programs including programs that cater to individual ethnic and racial groups such as black, white, Hispanic, and others. [George and Waldfoegel \(2003\)](#) obtained analogous results for newspapers and related readership. These and other studies make clear that larger metropolitan areas offer a more diverse and often a more appealing mix of consumption opportunities, but not necessarily in a manner that would raise or lower indicators of *per capita* income in the area.

Reinforcing this view, [Eeckhout et al. \(2014\)](#) develop a model of spatial sorting with skill complementarity in production. In their model, large metropolitan areas boost the productivity of *both* high-skill and very low-skill workers in what they refer to as “extreme skill complementarity.” This occurs because low-skill service workers provide valuable services that free high-skilled individuals to devote more time to market work. Empirically, Eeckhout, Pinheiro, and Schmidheiny find that both the highest- and the lowest-skill workers are overrepresented among the largest US cities and that both groups enjoy a skill-adjusted city size wage premium.⁵⁶ The endogenous growth of both high- and low-skill populations in larger metropolitan areas contributes to the diversity of consumption opportunities across the economic spectrum and further obscures any *a priori* tendency for large-city amenities to elevate *per capita* income and the city’s economic status.

The conclusion above is based implicitly on an assumption that high-amenity metropolitan areas are not in limited supply. [Gyourko et al. \(2013\)](#) considered a set of “superstar” cities with exceptionally attractive amenities for which that is not the case. San Francisco and Vancouver, BC, with their dramatic natural beauty are examples as is New York, which has no match in North America for its depth of cultural amenities including Broadway Theater and the like. Gyourko, Mayer, and Sinai argued that as national income rises, demand for residential opportunities in Superstar cities like San Francisco increases causing housing prices to shoot up. The increase in home prices discourages entry of low-income families into such areas and contributes to a high-income equilibrium. [Gyourko et al.](#)

⁵⁶ See [Baum-Snow and Pavan \(2012\)](#) for evidence of an alternative empirical pattern. They found that between 1979 and 2007, the largest cities in the United States had the largest growth in wage inequality. They attributed a substantial portion of the relative increase in large-city wage inequality to changes in the price for unobservable skills and lower relative wages for the lowest-skill workers.

(2013) provided empirical support for their model. They showed that as the number of high-income households in the United States grows, housing prices in high-amenity cities grow faster than in other cities and both average income and the upper tail of the income distribution in Superstar cities increase relative to other metropolitan areas.

As national incomes rise over time, an implication of the Superstar city model is that there could be a greater divergence in *per capita* income between high- and low-amenity metropolitan areas. [Eeckhout et al. \(2014\)](#), however, remind us that high- and low-skill individuals often complement each other, and for that reason, a growing population of higher-income individuals does not necessarily mean that lower-income households will be excluded from high-amenity metropolitan areas.

16.3.4.3 Inelastic housing supply

For growing cities with elastic housing supply, developers will expand the housing stock to meet demand. Under such conditions, housing supply would not have any long-term effects on MSA economic status and related dynamics. If instead, however, zoning limits new construction, then new housing supply will be inelastic and rising demand for the location is likely to be associated with rising house prices. If a metropolitan area is shrinking, the durability of the housing stock also causes housing supply to be highly inelastic in the face of declining demand. We explore implications of each of these scenarios for metropolitan income dynamics below.

16.3.4.3.1 Regulation in growing cities

An important set of recent studies demonstrates that regulation in growing, densely developed cities has reduced housing construction and caused housing rents and prices to rise. [Glaeser and Ward \(2009\)](#) showed that land use regulations in Massachusetts are associated with reduced levels of new construction activity and higher housing prices. They also showed that the effect of regulation on prices operates entirely through reductions in population density. [Glaeser et al. \(2005a,b\)](#) showed that housing prices are much higher than production costs in large, growing metropolitan areas like New York and argued that land use regulations are the natural explanation for this difference. For a sample of 44 US metropolitan areas, [Mayer and Tsuriel Somerville \(2000\)](#) reported that land use regulations are associated with lower levels of new construction and lower housing supply price elasticities. In California, [Quigley and Raphael \(2005\)](#) found that growth restrictions in heavily regulated cities lead to higher rents and housing prices, significantly reduced construction activity, and near-zero housing supply elasticities.

As a group, these studies suggest that housing regulations in growing cities can cause the supply of new housing to be extremely inelastic. In such locations, rising demand for residential opportunities should push price up. The effect of housing regulations on a metropolitan area's level of *per capita* income, however, is not completely obvious as this will depend on the impetus for rising demand and also the degree to which housing

regulations are associated with enhanced amenities that are favored by the rich. A prominent set of European cities, for example, have unusually attractive historic centers that are filled with cultural amenities, including Paris (as discussed by [Brueckner et al., 1999](#)), Barcelona, Seville, Amsterdam, Vienna, and more. To the extent that such amenities are of sufficient scale to attract higher-income households to the metro area, related regulations that restrict density and preserve important cultural sites have potential to elevate a city to superstar status with the associated rising income profile discussed by [Gyourko et al. \(2013\)](#). On the other hand, if regulations and related increases in housing costs are not matched by enhanced local amenities, then zoning could discourage migration into a metropolitan area. In this instance, labor supply could be reduced with potentially adverse effects on a city's *per capita* income and economic status.⁵⁷

16.3.4.3.2 Durable housing in declining cities

As emphasized in [Section 16.1](#), the US rust belt led by Detroit is a graphic reminder that cities can shrink.⁵⁸ [Glaeser and Gyourko \(2005\)](#) emphasized that in such instances, the durability of housing stock ensures that supply is highly inelastic, which contributes to sharp declines in house prices in the face of declining demand. As price falls below construction cost, new construction stops, which further reduces employment and demand, resulting in still lower home values. The resulting low home prices serve as a net attractor of low-skill, low-income families and other households not closely tied to the workforce (e.g., retirees). As the skill level in the city drops, it becomes even less attractive to business owners in search of skilled labor, encouraging still more businesses to leave the town. Durable housing, therefore, accelerates and deepens economic decline in shrinking cities.

[Glaeser and Gyourko \(2005\)](#) followed US MSAs over the 1970–2000 period and reported a series of different patterns that are all consistent with the model just described. Among other things, they documented that low-skill workers are attracted to the low housing prices in declining cities. An implication is that the durable nature of housing contributes to decline in *per capita* income and economic status in shrinking cities.⁵⁹

⁵⁷ In related work, [Saks \(2008\)](#) showed that as regulation levels increase, positive demand shocks have a reduced impact on employment growth but an enhanced positive impact on wage.

⁵⁸ It is worth noting that patterns reported in [Black and Henderson \(2003\)](#) indicate that the presence of major declining cities in the United States is a relatively recent phenomena.

⁵⁹ Extending these arguments, [Broxterman and Yezer \(2014\)](#) note that if the income elasticity of housing demand is less than one, then high-skill/high-earnings potential workers will tend to sort into high-rent, high-wage cities and low-skill workers will sort into low-rent, low-wage cities. The documented low-income elasticity of demand for housing, therefore, contributes to net migration of high-skill individuals away from declining metropolitan areas and reinforces the decline in such locations. [Yoon \(2013\)](#) also suggested that high-skill workers in formerly thriving cities like Detroit were more likely to have invested in skills that are readily marketable in other cities and that facilitate their out-migration during the city's decline. Low-skill workers, in contrast, were relatively more likely to have developed industry-specific skills that are more tied to a given city and less marketable elsewhere. Yoon argued that these differences reinforce decline in places such as Detroit.

Options for how to turn the downward spiral around are limited. Nevertheless, in recognition that the housing stock becomes a liability in declining cities, numerous US rust belt cities have begun to bulldoze housing stock. The goal in such policies is to shrink the supply of housing, boost prices, and break the city's downward spiral. Indeed, in 2009, Mayor Dave Bing in Detroit proposed bulldozing up to one-quarter of Detroit's housing stock!⁶⁰

Research discussed in earlier sections suggests that metropolitan areas typically exhibit parallel long-run growth trends, relatively static population ranks, and moderate churning of industrial composition. Our own analysis in Table 16.6 suggests comparatively rapid change in the economic status of a metropolitan area as measured by various indicators of *per capita* income. For all four of the stylized facts outlined at the start of this section, however, it is not clear that those patterns would hold equally for declining versus growing regions. In part, that is because the Glaeser–Gyourko model of durable housing makes clear that history should matter much more for declining metropolitan areas as compared to growing areas.

16.4. CONCLUSIONS AND FUTURE RESEARCH

A theme throughout this review is that the location of high- and low-income communities at both the neighborhood and metropolitan level arises from a blend of static and dynamic forces. In some instances, systematic cycling of a community's economic status should be anticipated. In other cases, change is driven more by random shocks that have potential to live on for extended periods of time. In still other instances, spatial patterns of development display incredible persistence.

The rise, fall, and renewal of Harlem is a reminder that communities can undergo dramatic change in economic status. The remarkable decline of Detroit also reminds us that just because a city is strong today, it will not necessarily always be so. In both cases, long periods of time are typically necessary to appreciate that a location's economic status may well be radically different in the future.

There remain numerous opportunities for further research on neighborhood and metropolitan income dynamics. Most obvious to us, at the neighborhood level, there

⁶⁰ See, for example, Snyder's (2010) article in *Business Insider*, "The Mayor of Detroit's Radical Plan to Bulldoze One Quarter of the City" at <http://www.businessinsider.com/the-mayor-of-detroits-radical-plan-to-bulldoze-one-quarter-of-the-city-2010-3>.

is a dearth of research on neighborhood income dynamics outside of a US context, including older European cities as well as emerging Asian giants. We also need to know more about whether gentrification hurts or helps existing residents since local opposition to gentrification and neighborhood change is so often tied to this issue. At the metropolitan level, there is need for further research on the nature and consequences of declining metropolitan areas. Although rising world population ensures that most cities will grow, shrinking cities are found in many places beyond the US rust belt including former manufacturing hubs in Europe, parts of Russia, and the former Soviet States (see, for example, [Power et al., 2008](#); [Berlinger, 2012](#); the [UN-Habitat's, 2013](#) report “State of the World’s Cities”). Research is also needed to help explain why metropolitan-level measures of *per capita* income exhibit more rapid change over time relative to industry churning and MSA population. Will growing superstar cities retain both low-skill labor and high-skill labor, and how will markets and/or government provide housing opportunities for lower-income families in such locations?

Finally, our review makes a distinction between neighborhood- and metropolitan-level income dynamics, mostly treating the two as independent events. Although this is a meaningful distinction, MSA-level events can nevertheless affect neighborhood sorting and vice versa. In a recent lecture, [Epple \(2012\)](#), for example, argued that decline in the Pittsburgh metropolitan area disproportionately hurt central city school districts and their communities. This occurs because declining metro-level population along with durable housing causes suburban house prices to fall, which draws central city residents to the suburbs in search of better schools and related amenities. The resulting hollowing out of a central city seems likely to be especially severe in declining metropolitan areas (see, for example, [Schmitt, 2010](#), “Without a Plan, Sprawl Will Continue to Hollow Out Cleveland Region”). In the opposite direction, [Benabou \(1996\)](#) developed a model in which neighborhood-level segregation of households by skill and financial resources has potential to undermine education production and human capital accumulation, depressing economic growth at the city level. [Fernandez and Rogerson \(1998\)](#) similarly argued that elimination of local funding disparities through school finance reform will yield large welfare gains that extend beyond immediate school districts. Related work on neighborhood based word-of-mouth job networks provides a further explanation for why the nature of household sorting at the neighborhood level can affect the quality of employer–employee job matches and labor productivity at the metropolitan level (e.g., [Damm, 2014](#); [Hellerstein et al., 2014](#); [Topa and Zenou, this volume](#)). Although these and other studies imply links between neighborhood- and metropolitan-level income dynamics, this also remains an area for further research.

Appendix Supplemental Figures

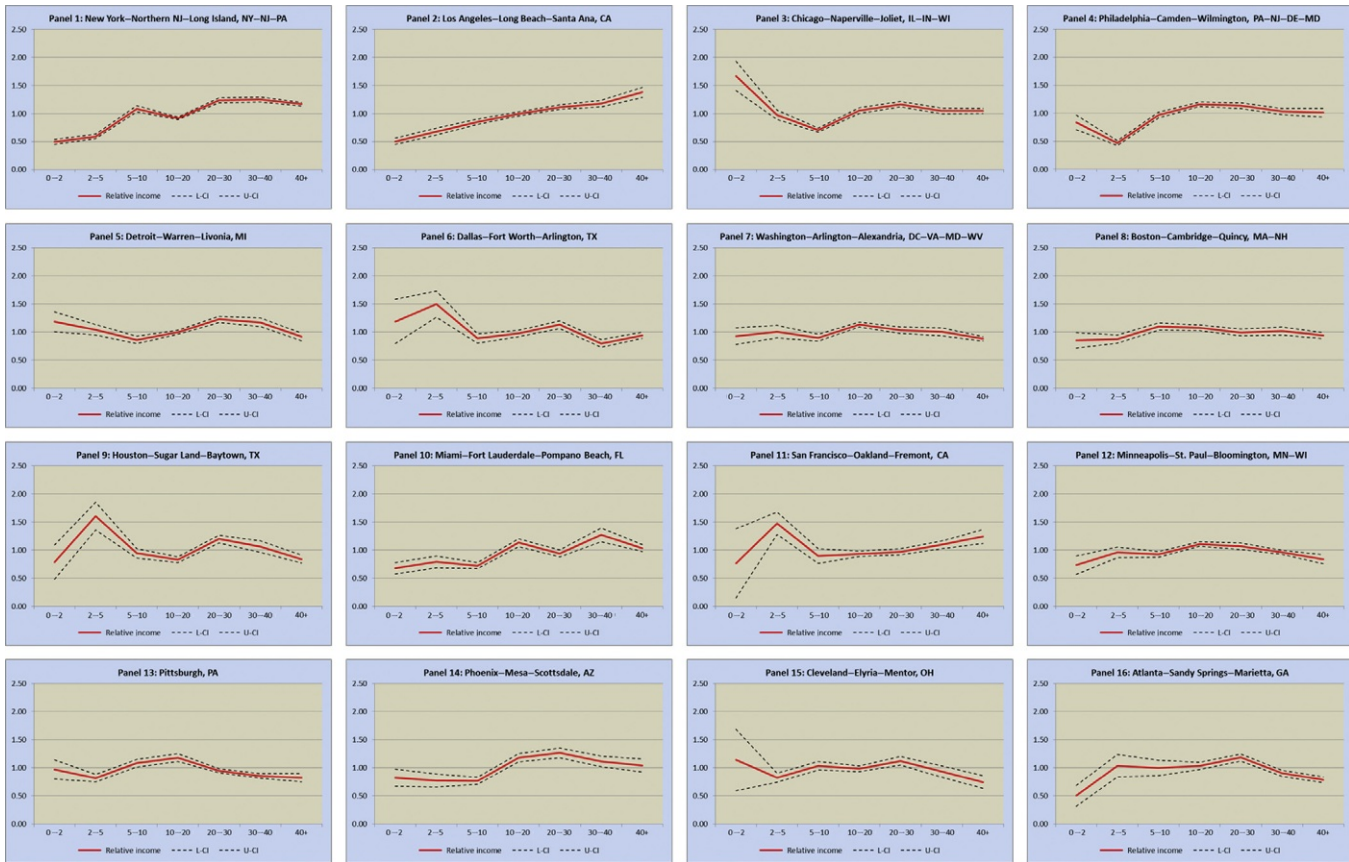


Figure A1 Census tract income relative to the MSA by distance (in miles) to the MSA center using the 2005–2009 ACS pooled sample.

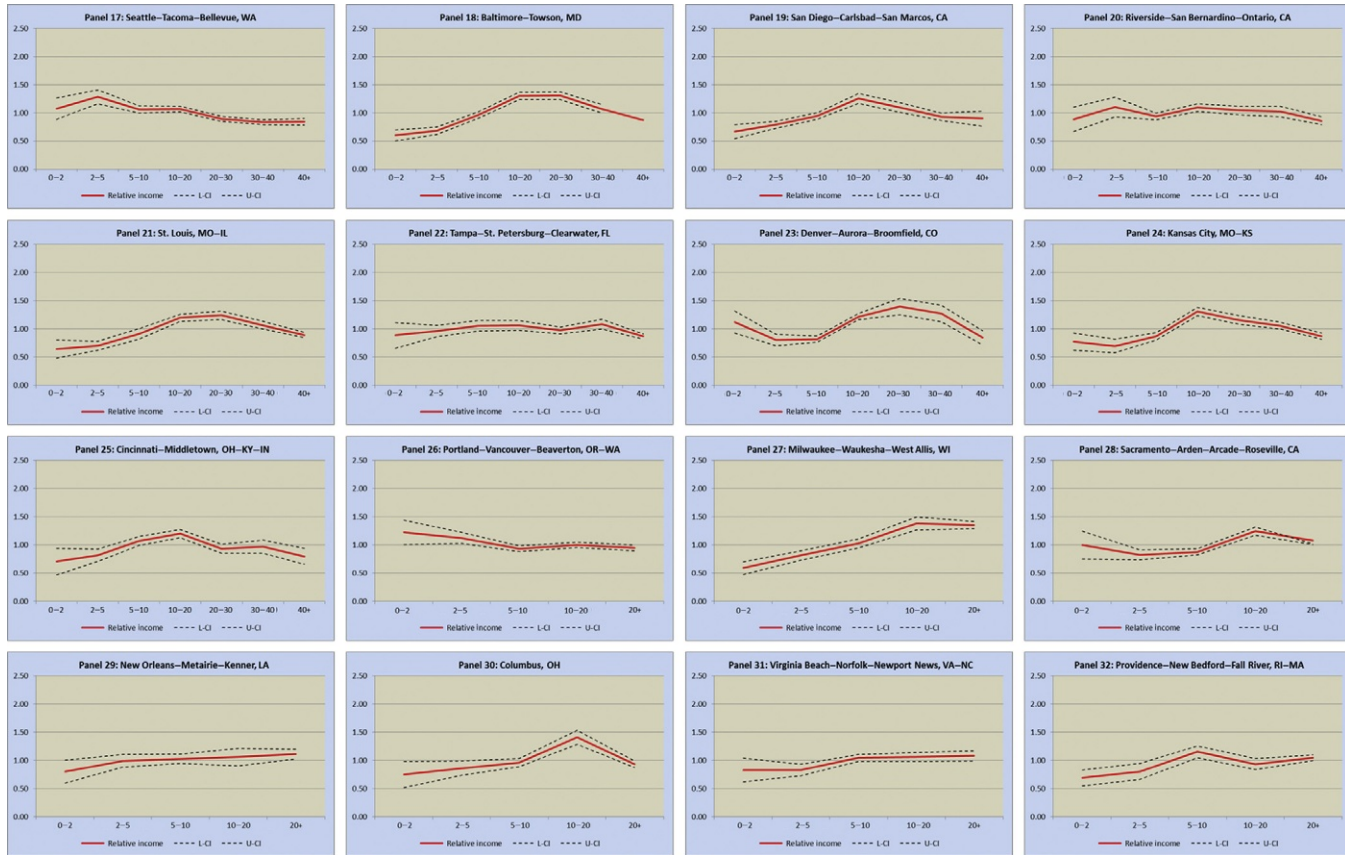


Figure A1—Cont'd

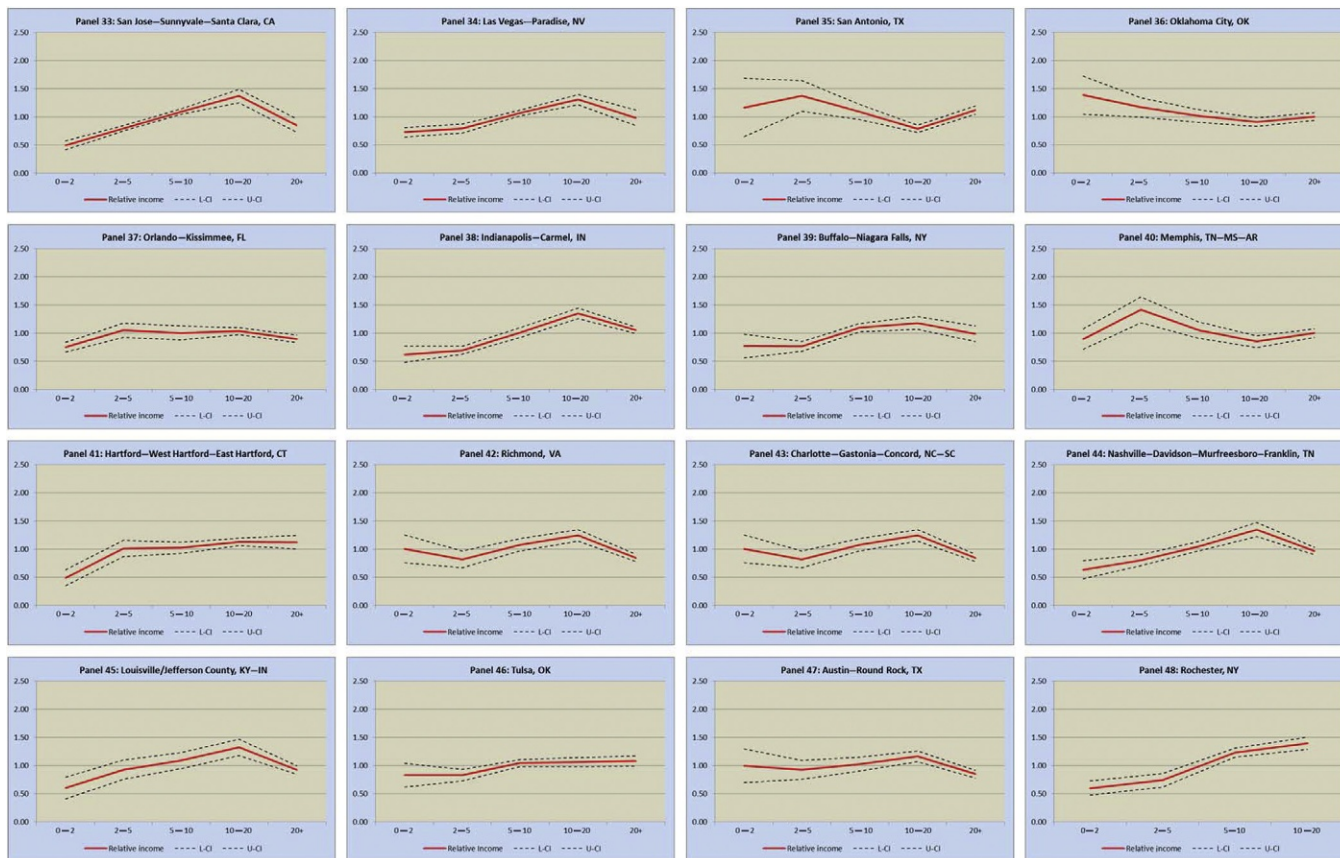


Figure A1—Cont'd

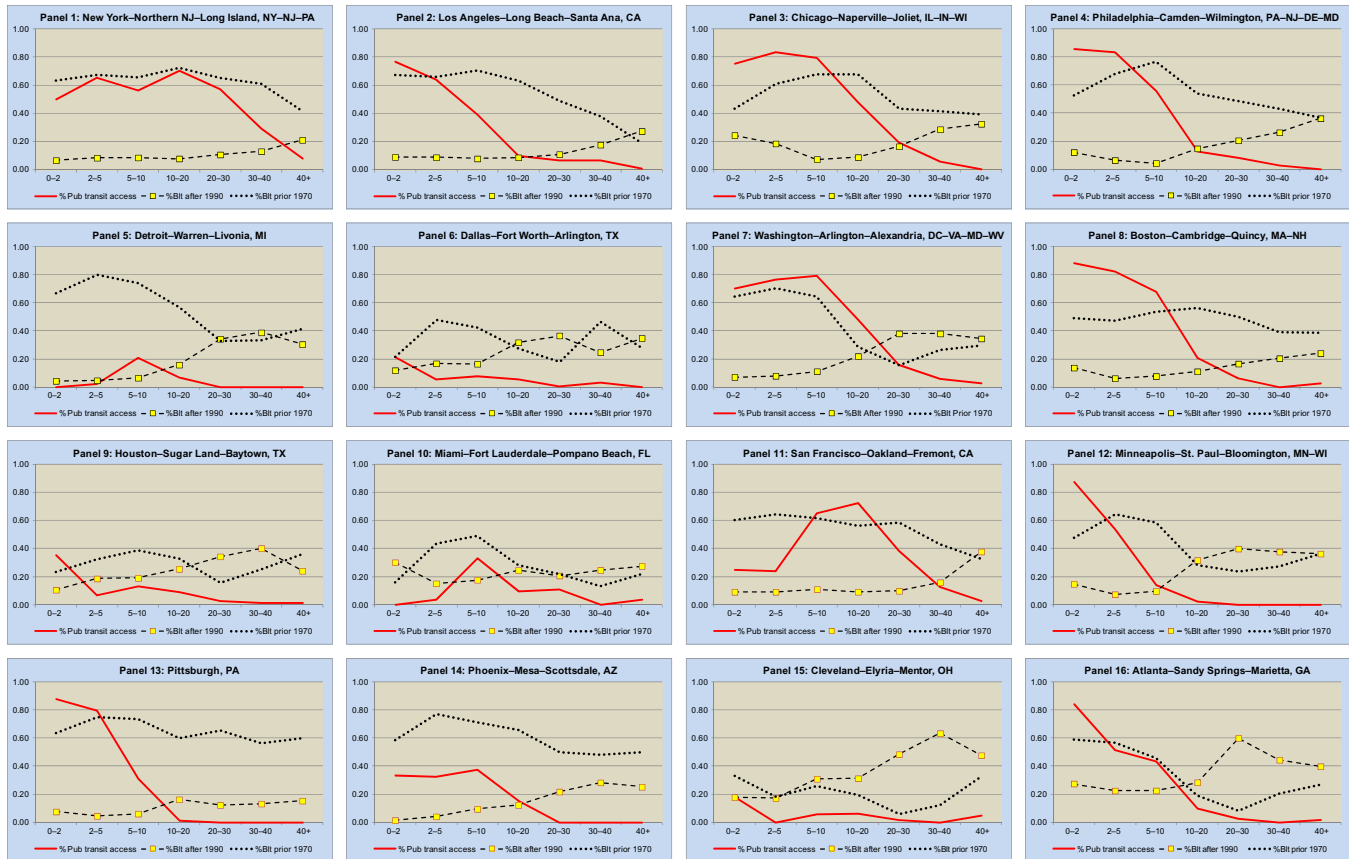


Figure A2 Convenient access to public transit for work trips and age of housing stock by census tract distance (in miles) to the MSA center using the 2005–2009 ACS pooled sample.

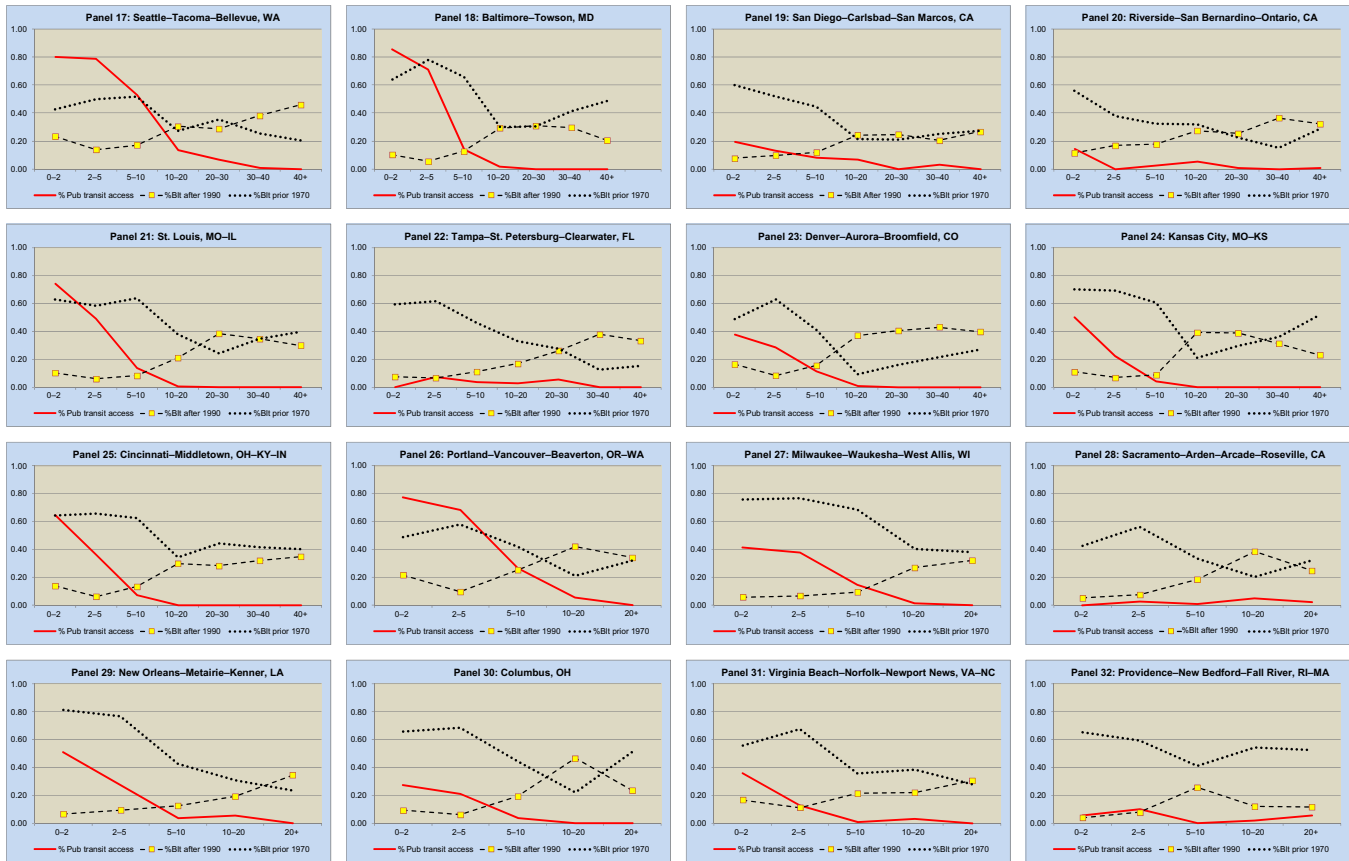


Figure A2—Cont'd

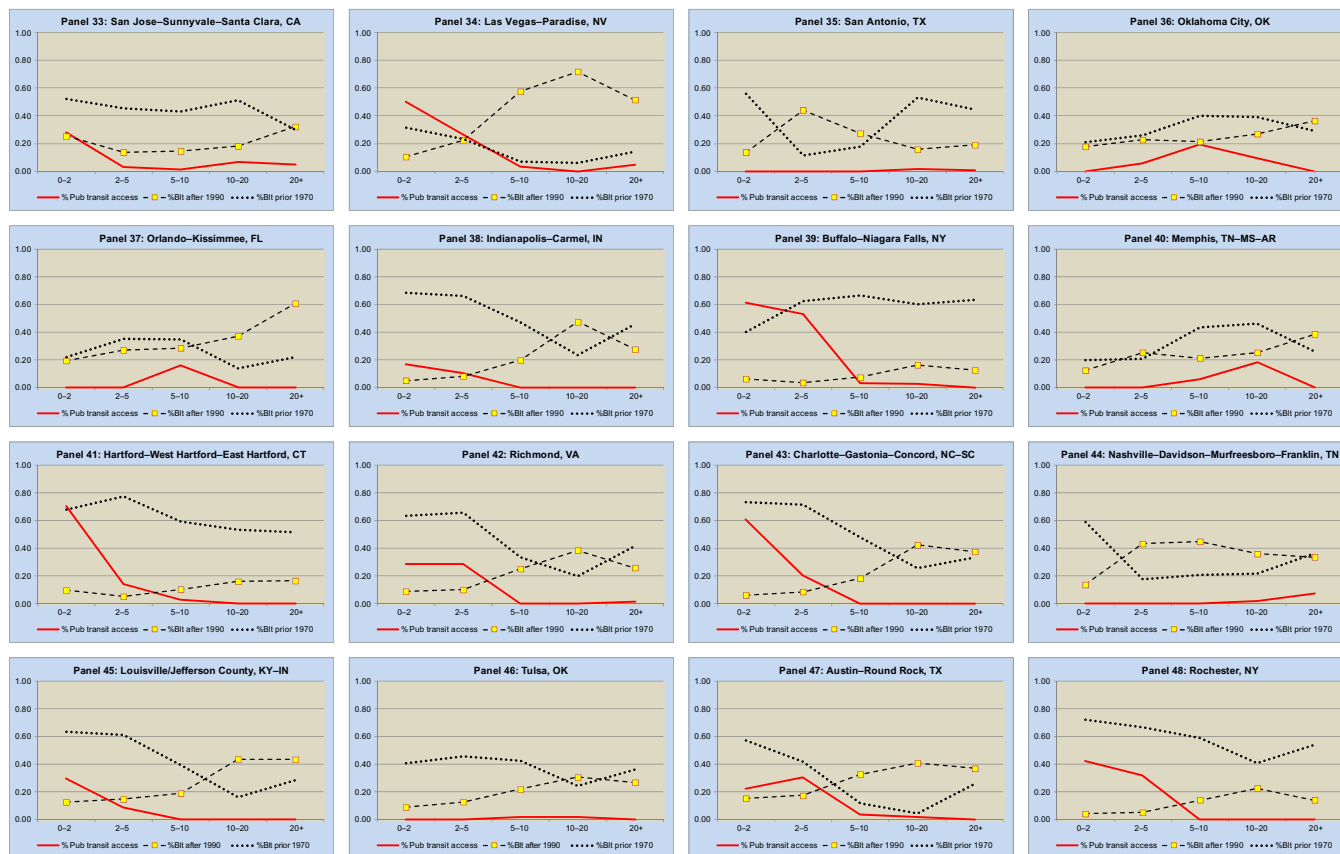


Figure A2—Cont'd

ACKNOWLEDGMENTS

We thank Matthew Turner, William Strange, Gilles Duranton, and Vernon Henderson for helpful comments. Nuno Mota provided excellent research assistance. The usual disclaimer applies.

REFERENCES

- Acemoglu, D., Johnson, S., Robinson, J., 2005. The rise of Europe: Atlantic trade, institutional change, and economic growth. *Am. Econ. Rev.* 95 (3), 546–579.
- Ahlfeldt, G.M., Maennig, W., 2014. Homevoters vs. leasevoters: a spatial analysis of airport effects. Working paper.
- Albouy, D., 2008. Are big cities bad places to live? Estimating quality of life across metropolitan areas: National Bureau of Economic Research. NBER Working paper No. 14472, November.
- Albouy, D., 2009. The unequal geographic burden of federal taxation. *J. Polit. Econ.* 117 (4), 635–667.
- Alesina, A., Baqir, R., Easterly, W., 1999. Public goods and ethnic divisions. *Q. J. Econ.* 114 (4), 1243–1284.
- Alonso, W., 1964. *Location and Land Use*. Harvard University Press, Cambridge.
- Arnott, R.J., Braid, R.M., 1997. A filtering model with steady-state housing. *Reg. Sci. Urban Econ.* 27 (4–5), 515–546.
- Banzhaf, H.S., Walsh, R.P., 2008. Do people vote with their feet? An empirical test of Tiebout's mechanism. *Am. Econ. Rev.* 98 (3), 843–863.
- Bar-Ilan, A., Strange, W.C., 1996. Urban development with lags. *J. Urban Econ.* 39 (10), 87–113.
- Baum-Snow, N., 2007. Did highways cause suburbanization? *Q. J. Econ.* 122 (2), 775–805.
- Baum-Snow, N., Pavan, R., 2012. Understanding the city size wage gap. *Rev. Econ. Stud.* 79 (1), 88–127.
- Bayer, P., McMillan, R., 2005. Racial sorting and neighborhood quality. NBER Working paper 11813, December. National Bureau of Economic Research.
- Bayer, P., McMillan, R., 2012. Tiebout sorting and neighborhood stratification. *J. Public Econ.* 96 (11), 1129–1143.
- Bayer, P., McMillan, R., Rueben, K., 2004. What drives racial segregation? New evidence using census microdata. *J. Urban Econ.* 56 (3), 514–535.
- Bayer, P., McMillan, R., Rueben, K., 2005. Residential segregation in general equilibrium. National Bureau of Economic Research. NBER Working paper No. 11095, January.
- Bayer, P., Fang, H., McMillan, R., 2014. Separate when equal? Racial inequality and residential segregation. *J. Urban Econ.* 82, 32–48.
- Behrens, K., Robert-Nicoud, F., this volume. Agglomeration theory. In: Duranton, G., Vernon Henderson, J., Strange, W.C. (Eds.), *Handbook of Regional and Urban Economics*, vol. 5.
- Benabou, R., 1993. Workings of a city: location, education, and production. *Q. J. Econ.* 108 (3), 619–652.
- Benabou, R., 1996. Heterogeneity, stratification, and growth: macroeconomic implications of community structure and school finance. *Am. Econ. Rev.* 86 (3), 584–609.
- Berlinger, J., 2012. The 28 fastest-shrinking cities in the world. *Business Insider* (October 22). <http://www.businessinsider.com/cities-that-are-shrinking-2012-10?op=1>.
- Black, D., Vernon Henderson, J., 2003. Urban evolution in the USA. *J. Econ. Geogr.* 3 (4), 343–372.
- Bleakley, H., Lin, J., 2012. Portage and path dependence. *Q. J. Econ.* 127, 587–644.
- Blomquist, G., Berger, M., Hoehn, J., 1988. New estimates of the quality of life in urban areas. *Am. Econ. Rev.* 78, 89–107.
- Boehm, T.P., Schlottmann, A.M., 2004. The dynamics of race, income, and homeownership. *J. Urban Econ.* 55 (1), 113–130.
- Bond, E.W., Coulson, N.E., 1989. Externalities, filtering, and neighborhood change. *J. Urban Econ.* 26 (2), 231–249.
- Boustan, L.P., 2010. Was postwar suburbanization 'white flight'? Evidence from the black migration. *Q. J. Econ.* 125 (1), 417–443.
- Boustan, L.P., Margo, R.A., 2013. A silver lining to white flight? White suburbanization and African-American homeownership, 1940–1980. *J. Urban Econ.* 74 (1), 71–80.

- Braid, R.M., 2001. Spatial growth and redevelopment with perfect foresight and durable housing. *J. Urban Econ.* 49 (3), 425–452.
- Brakman, S., Garretsen, H., Schramm, M., 2004. The strategic bombing of German cities during World War II and its impact on city growth. *J. Econ. Geogr.* 4, 201–218.
- Brooks, L., Lutz, B., 2012. From today's city to tomorrow's city: an empirical investigation of urban land assembly: Federal Reserve Board of Governors. Working paper, November.
- Broxterman, D.A., Yezer, A.M., 2014. City size and skill intensity: is it all housing cost? Social Science Research Network Working paper, SSRN-id2408179.
- Brueckner, J., 2011. *Lectures on Urban Economics*. The MIT Press, Cambridge.
- Brueckner, J., Helsley, R., 2011. Sprawl and blight. *J. Urban Econ.* 69 (1), 205–213.
- Brueckner, J., Rosenthal, S.S., 2009. Gentrification and neighborhood cycles: will America's future down-towns be rich? *Rev. Econ. Stat.* 91 (4), 725–743.
- Brueckner, J., Thisse, J.F., Zenou, Y., 1999. Why is central Paris rich and downtown Detroit poor? An amenity-based theory. *Eur. Econ. Rev.* 43, 91–107.
- Brunner, E., 2014. School quality, school choice, and residential mobility. In: Ingram, G.K., Kenyon, D.A. (Eds.), *Education, Land, and Location*. Lincoln Institute of Land Policy, Cambridge, MA. http://www.lincolninstitute.edu/pubs/2398_School-Quality-School-Choice-and-Residential-Mobility.
- Brunner, E., Sonstelie, J., 2003. Homeowners, property values, and the political economy of the school voucher. *J. Urban Econ.* 54, 239–257.
- Brunner, E., Sonstelie, J., Thayer, M., 2001. Capitalization and the voucher: an analysis of precinct returns from California's proposition 174. *J. Urban Econ.* 50, 517–536.
- Bulan, L., Mayer, C., Tsuriel Somerville, C., 2009. Irreversible investment, real options, and competition: evidence from real estate development. *J. Urban Econ.* 65 (3), 237–251.
- Calabrese, S., Epple, D., Romer, T., Sieg, H., 2006. Local public good provision: voting, peer effects, and mobility. *J. Public Econ.* 90 (6), 959–981.
- Capozza, D.R., Helsley, R.W., 1990. The stochastic city. *J. Urban Econ.* 28 (2), 187–203.
- Card, D., Mas, A., Rothstein, J., 2008. Tipping and the dynamics of segregation. *Q. J. Econ.* 123 (1), 177–218.
- Card, D., Mas, A., Rothstein, J., 2010. Are mixed neighborhoods always unstable? Two-sided and one-sided tipping. In: Newburger, H., Birch, E.L., Wachter, S.M. (Eds.), *Neighborhood and Life Changes: How Place Matters in Modern America?* University of Pennsylvania Press, Philadelphia, PA, pp. 237–256.
- Carlino, G.A., Kerr, W., this volume. Agglomeration and innovation. In: Duranton, G., Vernon Henderson, J., Strange, W.C. (Eds.), *Handbook of Regional and Urban Economics*, vol. 5.
- Case, K.E., Shiller, R.J., 1987. Prices of Single Family Homes Since 1970: New Indexes for Four Cities. *New England Economic Review*, Boston, MA, pp. 45–56.
- Chandler, T., 1987. *Four Thousand Years of Urban Growth: An Historical Census*. The Edwin Mellen Press, Lewiston, NY.
- Charles, K.K., Hurst, E., 2002. The transition to home ownership and the black-white wealth gap. *Rev. Econ. Stat.* 84 (2), 281–297.
- Chen, Y., Rosenthal, S.S., 2008. Local amenities and life cycle migration: do people move for jobs or fun? *J. Urban Econ.* 65 (3), 519–537.
- Choi, C.Y., 2004. A reexamination of output convergence in the U.S. states: toward which level(s) are they converging? *J. Reg. Sci.* 44 (4), 713–741.
- Christaller, W., 1966. *Central Places in Southern Germany* (Carlisle W. Baskin, Trans.). Prentice Hall, Inc., Englewood Cliffs, NJ.
- Clapp, J.M., Salavei, K., 2010. Hedonic pricing with redevelopment options: a new approach to estimating depreciation effects. *J. Urban Econ.* 67 (3), 362–377.
- Combes, P.P., Gobillon, L., this volume. The empirics of agglomeration. In: Duranton, G., Vernon Henderson, J., Strange, W.C. (Eds.), *Handbook of Regional and Urban Economics*, vol. 5.
- Corominas-Murtra, B., Hanel, R., Thurner, S., 2014. Understanding Zipf's Law with Playing Dice: History-Dependent Stochastic Processes with Collapsing Sample-Space Have Power-Law Rank Distributions. Cornell University Library, Ithaca, NY, arXiv: 1407.2775v2, 12 July 2014.
- Courant, P.N., Yinger, J., 1977. On models of racial prejudice and urban residential structure. *J. Urban Econ.* 4 (3), 272–291.

- Cuberes, D., 2011. Sequential city growth: empirical evidence. *J. Urban Econ.* 69 (2), 229–239.
- Cutler, D., Glaeser, E., Vigdor, J., 1999. The rise and decline of the American ghetto. *J. Polit. Econ.* 107 (3), 455–506.
- Damm, A., 2014. Neighborhood quality and labor market outcomes: evidence from quasi-random neighborhood assignment of immigrants. *J. Urban Econ.* 79, 139–166.
- Davidson, J., 2014. Is gentrification all bad? *New York Magazine* (February 2). <http://nymag.com/news/features/gentrification-2014-2/>.
- Davis, D.R., Weinstein, D.E., 2002. Bones, bombs, and breakpoints: the geography of economic activity. *Am. Econ. Rev.* 92 (5), 1269–1289.
- Davis, D.R., Weinstein, D.E., 2008. A search for multiple equilibria in Urban industrial structure. *J. Reg. Sci.* 48 (1), 29–65.
- Dawkins, C.J., 2005. Racial gaps in the transition to first-time homeownership: the role of residential location. *J. Urban Econ.* 58 (3), 537–554.
- DeBartolomé, C.A.M., 1990. Equilibrium and inefficiency in a community model with peer group effects. *J. Polit. Econ.* 98 (1), 110–133.
- DeBartolomé, C.A.M., Ross, S.L., 2003. Equilibria with local governments and commuting: income sorting vs income mixing. *J. Urban Econ.* 54 (1), 1–20.
- DeBartolomé, C.A.M., Ross, S.L., 2008. The race to the suburb: the location of the poor in a metropolitan area. University of Connecticut, Department of Economics Working paper, May.
- Dehring, C.A., Depken II, C.A., Ward, M.R., 2008. A direct test of the homevoter hypothesis. *J. Urban Econ.* 64, 155–170.
- Deng, Y., Ross, S.L., Wachter, S.M., 2003. Racial differences in homeownership: the effect of residential location. *Reg. Sci. Urban Econ.* 33 (5), 517–556.
- Desmet, K., Henderson, V., this volume. Cities and development. In: Duranton, G., Vernon Henderson, J., Strange, W.C. (Eds.), *Handbook of Regional and Urban Economics*, vol. 5.
- Desmet, K., Rappaport, J., 2013. The settlement of the United States, 1800 to 2000: the long transition towards Gibrat's law: Federal Reserve Bank of Kansas City. Research Working paper 13–02, September.
- Duranton, G., 2007. Urban evolutions: the fast, the slow, and the still. *Am. Econ. Rev.* 97 (1), 197–221.
- Duranton, G., Puga, D., 2001. Nursery cities: urban diversity, process innovation, and the life cycle of products. *Am. Econ. Rev.* 91 (5), 1454–1477.
- Duranton, G., Puga, D., 2004. Micro-foundations of urban agglomeration economies. In: Henderson, V., Thisse, J. (Eds.), *Handbook of Urban and Regional Economics*, vol. 4. Elsevier, Amsterdam, Netherlands, pp. 2063–2117.
- Dye, R.F., McMillen, D.P., 2007. Teardowns and land values in the Chicago metropolitan area. *J. Urban Econ.* 61 (1), 45–63.
- Eaton, J., Eckstein, Z., 1997. Cities and growth: theory and evidence from France and Japan. *Reg. Sci. Urban Econ.* 27 (4), 443–447.
- Eeckhout, J., Pinheiro, R., Schmidheiny, K., 2014. Spatial sorting. *J. Polit. Econ.* 122 (3), 554–620.
- Epplé, D., 2012. School choice and urban school districts. Distinguished Lecture in Economics, University of Connecticut, April 12.
- Epplé, D., Ferreyra, M.M., 2008. School finance reform: assessing general equilibrium effects. *J. Public Econ.* 92 (5), 1326–1351.
- Epplé, D., Romano, R.E., 1998. Competition between private and public schools, vouchers, and peer-group effects. *Am. Econ. Rev.* 88 (1), 33–62.
- Epplé, D., Romano, R.E., 2008. Educational vouchers and cream skimming. *Int. Econ. Rev.* 49 (4), 1395–1435.
- Epplé, D., Romano, R.E., 2011. Peer effects in education: a survey of the theory and evidence. In: Benhabib, J., Bisin, A., Jackson, M.O. (Eds.), *Handbook of Social Economics*. Elsevier, Amsterdam, Netherlands, pp. 1053–1163 (Chapter 20).
- Epplé, D., Romer, T., 1991. Mobility and redistribution. *J. Polit. Econ.* 99 (4), 828–858.
- Epplé, D., Gordon, B., Sieg, H., 2010. A new approach to estimating the production function for housing. *Am. Econ. Rev.* 100 (3), 905–924.

- Falck, O., Heblich, S., Lameli, A., Südekum, J., 2012. Dialects, cultural identity, and economic exchange. *J. Urban Econ.* 72 (2), 225–239.
- Fernandez, R., Rogerson, R., 1998. Income distribution and public education: a dynamic quantitative evaluation of school finance reform. *Am. Econ. Rev.* 88, 813–833.
- Ferreira, M.M., 2007. Estimating the effects of private school vouchers in multidistrict economies. *Am. Econ. Rev.* 97 (3), 789–817.
- Ferreira, M.M., 2009. An empirical framework for large-scale policy analysis, with an application to school finance reform in Michigan. *Am. Econ. J. Econ. Policy* 1 (1), 147–180.
- Findeisen, S., Südekum, J., 2008. Industry churning and the evolution of cities: evidence for Germany. *J. Urban Econ.* 64 (2), 326–339.
- Fischel, F.A., 2001. *The Homevoter Hypothesis*. Harvard University Press, Cambridge.
- Freeman, L., 2005. Displacement or succession? Residential mobility in gentrifying neighborhoods. *Urban Aff. Rev.* 40 (4), 463–491.
- Gabaix, X., Ioannides, Y.M., 2004. The evolution of city size distributions. In: Vernon Henderson, J., Thisse, J.-F. (Eds.), *Handbook of Regional and Urban Economics*, vol. 4. Elsevier, Amsterdam, Netherlands, pp. 2341–2378 (Chapter 53).
- Gabriel, S.A., Rosenthal, S.S., 2005. Homeownership in the 1980s and 1990s: aggregate trends and racial gaps. *J. Urban Econ.* 57 (1), 101–127.
- George, L., Waldfogel, J., 2003. Who affects whom in daily newspaper markets? *J. Polit. Econ.* 111 (4), 765–784.
- Glaeser, E.L., 2007. Can buffalo ever come back? *City J.* (Autumn).
- Glaeser, E.L., Gyourko, J., 2005. Urban decline and durable housing. *J. Polit. Econ.* 113 (2), 345–375.
- Glaeser, E.L., Kahn, M., 2001. Decentralized employment and the transformation of the American city. *Brookings-Wharton Papers on Urban Affairs*, vol. 2. Brookings Institution Press, Washington, DC, pp. 1–63.
- Glaeser, E.L., Mare, D.C., 2001. Cities and skills. *J. Labor Econ.* 19 (2), 316–342.
- Glaeser, E.L., Ward, B.A., 2009. The causes and consequences of land use regulation: evidence from Greater Boston. *J. Urban Econ.* 65 (3), 265–278.
- Glaeser, E.L., Sacerdote, B., Scheinkman, J.A., 1996. Crime and social interactions. *Q. J. Econ.* 111 (2), 507–548.
- Glaeser, E.L., Kolko, J., Saiz, A., 2001. Consumer city. *J. Econ. Geogr.* 1 (1), 27–50.
- Glaeser, E.L., Gyourko, J., Saks, R.E., 2005a. Why is Manhattan so expensive? Regulation and the rise in housing prices. *J. Law Econ.* XLVIII, 331–369.
- Glaeser, E.L., Gyourko, J., Saks, R.E., 2005b. Why have housing prices gone up? *Am. Econ. Rev.* 95 (2), 329–333.
- Glaeser, E.L., Kahn, M.E., Rappaport, J., 2008. Why do the poor live in cities? The role of public transportation. *J. Urban Econ.* 63 (1), 1–24.
- Glocker, D., Sturm, D.M., 2013. War-time destruction and the persistence of economic activity. London School of Economics. Working paper, London, November.
- Grossman, G.M., Helpman, E., 1991. Quality ladders in the theory of growth. *Rev. Econ. Stud.* 58 (1), 43–61.
- Guerrieri, V., Hartley, D., Hurst, E., 2013. Endogenous gentrification and housing price dynamics. *J. Public Econ.* 100, 45–60.
- Guthrie, G., 2010. House prices, development costs, and the value of waiting. *J. Urban Econ.* 68 (1), 56–71.
- Gyourko, J., Saiz, A., 2004. Reinvestment in the housing stock: the role of construction costs and the supply side. *J. Urban Econ.* 55 (2), 238–256.
- Gyourko, J., Tracy, J., 1991. The structure of local public finance and the quality of life. *J. Polit. Econ.* 99, 774–806.
- Gyourko, J., Kahn, M., Tracy, J., 1999. Quality of life and environmental comparisons. In: Mills, E.S., Cheshire, P. (Eds.), *Handbook of Urban and Regional Economics*, vol. 3. Elsevier, Amsterdam, Netherlands, pp. 1414–1443.
- Gyourko, J., Mayer, C., Sinai, T., 2013. Superstar cities. *Am. Econ. J. Econ. Policy* 5 (4), 167–199.

- Hanlon, W.W., Miscio, A., 2013. Agglomeration: a dynamic approach. UCLA Economics Department. Working paper, December.
- Hanushek, E., Yilmaz, K., 2007. The complementarity of Tiebout and Alonso. *J. Hous. Econ.* 16 (2), 243–261.
- Harding, J., Rosenthal, S.S., Sirmans, C.F., 2007. Depreciation of housing capital, maintenance, and house price inflation: estimates from a repeat sales model. *J. Urban Econ.* 61 (2), 193–217.
- Haurin, D.R., Herbert, C.E., Rosenthal, S.S., 2007. Homeownership gaps among low-income and minority households. *Citiescape* 9 (2), 5–52.
- Hellerstein, J., Kutzbach, M., Neumark, D., 2014. Do labor market networks have an important spatial dimension? *J. Urban Econ.* 79, 39–58.
- Hilber, C., Mayer, C., 2009. Why do households without children support local public schools? Linking house price capitalization to school spending. *J. Urban Econ.* 65 (1), 74–90.
- History of Harlem Heritage Tours & Cultural Center. Harlem Heritage Tours Cultural Center RSS. <http://www.harlemheritage.com/history-of-harlem/>.
- Holmes, M.J., Otero, J., Panagiotidis, T., 2013. A note on the extent of U.S. regional income convergence. *Macroecon. Dyn.* 17 (1), 1–28.
- Jacobs, J., 1969. *The Economy of Cities*. Vintage, New York.
- Jacobs, J., 1984. *Cities and the Wealth of Nations: Principles of Economic Life*. Vintage, New York.
- Jedwab, R., Moradi, A., 2014. Transportation technology and economic change: the impact of colonial railroads on city growth in Africa. The George Washington University, Institute for International Economic Policy. Working papers 2014–03, January.
- Jedwab, R., Kerby, E., Moradi, A., 2014. History, path dependence and development: evidence from colonial railroads, settlers and cities in Kenya. Centre for the Study of African Economies, University of Oxford. CSAE Working paper Series 2014–04, January.
- Lee, S., Lin, J., 2013. Natural amenities, neighborhood dynamics, and persistence in the spatial distribution of income: Federal Reserve Bank of Philadelphia Working paper 13–48.
- Lee, S., Ries, J., Tsurriel Somerville, C., 2013. Repairs under imperfect information. *J. Urban Econ.* 73 (1), 43–56.
- LeRoy, S.F., Sonstelie, J., 1983. Paradise lost and regained: transportation innovation, income, and residential location. *J. Urban Econ.* 13 (1), 67–89.
- Levin, A., Lin, C.-F., Chu, C.-S., 2002. Unit root test in panel data: asymptotic and finite sample properties. *J. Econ.* 108 (1), 1–25.
- Losch, A., 1954. *The Economics of Location*. Yale University Press, New Haven.
- Maddala, G.S., Wu, S., 1999. A comparative study of unit root tests with panel data and a new simple test. *Oxf. Bull. Econ. Stat.* 61 (S1), 631–652.
- Massey, D.S., Fischer, M.J., 2003. The geography of inequality in the United States, 1950–2000. *Brookings-Wharton Papers on Urban Affairs*, vol. 4. Brookings Institution Press, Washington, DC, pp. 1–40.
- Mayer, C.J., Tsurriel Somerville, C., 2000. Land use regulation and new construction. *Reg. Sci. Urban Econ.* 30 (6), 639–662.
- McKinnish, T., Walsh, R., Kirk White, T., 2010. Who gentrifies low-income neighborhoods? *J. Urban Econ.* 67 (1), 180–193.
- McMillan, R., 2005. Erratum to competition, incentives, and public school productivity. *J. Public Econ.* 89, 1133–1154.
- McMillen, D., O'Sullivan, A., 2013. Option value and the price of teardown properties. *J. Urban Econ.* 74 (1), 71–82.
- Mello, M., 2011. Stochastic convergence across U.S. states. *Macroecon. Dyn.* 15 (2), 160–183.
- Mills, E.S., 1967. An aggregative model of resource allocation in a metropolitan area. *Am. Econ. Rev.* 57, 197–210.
- Muth, R.F., 1969. *Cities and Housing*. University of Chicago Press, Chicago.
- Nurse, A., 2008. The Liverpool View: Detroit's Lessons for Liverpool. University of Liverpool, University News, Liverpool, England. <http://news.liv.ac.uk/2013/08/02/the-liverpool-view-detroits-lessons-for-liverpool/>.
- O'Sullivan, A., 2003. *Urban Economics*. McGraw-Hill Companies, Inc., New York, NY.
- O'Sullivan, A., 2005. Gentrification and crime. *J. Urban Econ.* 57 (1), 73–85.

- O'Sullivan, A., 2009. Schelling's model revisited: residential sorting with competitive bidding for land. *J. Urban Econ.* 39 (4), 397–408.
- Pancs, R., Vriend, N.J., 2007. Schelling's spatial proximity model of segregation revisited. *J. Public Econ.* 91, 1–24.
- Payne, L., 2010. Is Harlem no longer black? *The Root*. http://www.theroot.com/articles/culture/2010/01/is_harlem_no_longer_black.html.
- Power, A., Ploger, J., Winkler, A., 2008. Transforming cities across Europe: an interim report on problems and progress: CASEreport 49. Centre for Analysis of Social Exclusion, London School of Economics and Political Science.
- Prasad, P.C., 1977. *Foreign Trade and Commerce in Ancient India*. Abhinav Publications, New Delhi, p. 90.
- Quigley, J.M., 1998. Urban density and economic growth. *J. Econ. Perspect.* 12 (2), 127–138.
- Quigley, J.M., Raphael, S., 2005. Regulation and the high cost of housing in California. *Am. Econ. Rev.* 95 (2), 323–328.
- Roback, J., 1982. Wages, rents, and the quality of life. *J. Polit. Econ.* 90 (6), 1257–1278.
- Roberts, S., 2010. No longer majority black, Harlem is in transition. *The New York Times* (5 January). http://www.nytimes.com/2010/01/06/nyregion/06harlem.html?pagewanted=all&_r=0.
- Rosenthal, S.S., 2008a. Old homes, externalities, and poor neighborhoods: a model of urban decline and renewal. *J. Urban Econ.* 63 (3), 816–840.
- Rosenthal, S.S., 2008b. Where poor renters live in our cities: dynamics and determinant. In: Retsinas, N., Belskey, E. (Eds.), *Revisiting Rental Housing: A National Policy Summit*. Brookings Press, Washington, DC, pp. 59–92.
- Rosenthal, S.S., 2014. Are private markets and filtering a viable source of low-income housing? Estimates from a 'repeat income' model. *Am. Econ. Rev.* 104 (2), 687–706.
- Rosenthal, S.S., Helsley, R., 1994. Redevelopment and the urban land price gradient. *J. Urban Econ.* 35 (2), 182–200.
- Rosenthal, S.S., Strange, W.C., 2004. Evidence on the nature and sources of agglomeration economies. In: Henderson, V., Thisse, J. (Eds.), *Handbook of Urban and Regional Economics*, vol. 4. Elsevier, Amsterdam, Netherlands, pp. 2119–2172.
- Rosenthal, S.S., Strange, W.C., 2008. The attenuation of human capital spillovers. *J. Urban Econ.* 64 (2), 373–389.
- Ross, S., 2011. Social interactions within cities: neighborhood environments and peer relationships. In: Brooks, N., Donaghy, K., Knapp, G. (Eds.), *Handbook of Urban Economics and Planning*. Oxford University Press, New York, NY, pp. 203–229 (Chapter 9).
- Ross, S., Yinger, J., 1999. Sorting and voting: a review of the literature on urban public finance. In: Mills, E.S., Cheshire, P. (Eds.), *Handbook of Regional and Urban Economics*, vol. 3. Elsevier, Amsterdam, Netherlands, pp. 2001–2060 (Chapter 47).
- Ross, S., Yinger, J., 2002. *Color of Credit: Mortgage Discrimination, Research Methods, and Fair Lending Enforcement*. MIT Press, Cambridge.
- Saiz, A., 2010. The geographic determinants of housing supply. *Q. J. Econ.* 125 (3), 1253–1296.
- Saks, R.E., 2008. Job creation and housing construction: constraints on metropolitan area employment growth. *J. Urban Econ.* 64 (1), 178–195.
- Schelling, T.C., 1971. Dynamic models of segregation. *J. Math. Sociol.* 1, 143–186.
- Schelling, T.C., 1978. *Micromotives and Macrobehavior*. Norton Press, New York.
- Schmitt, A., 2010. Without a plan, Sprawl will continue to hollow out cleveland region. *StreetsBlog USA* (Monday, September 13). <http://usa.streetsblog.org/2010/09/13/without-a-plan-sprawl-will-continue-to-hollow-out-cleveland-region/>.
- Sharma, S., 2003. Persistence and stability in city growth. *J. Urban Econ.* 53 (2), 300–320.
- Snyder, M., 2010. The Mayor of Detroit's Radical Plan to Bulldoze One Quarter of the City. *Business Insider* (March 10). <http://www.businessinsider.com/the-mayor-of-detroits-radical-plan-to-bulldoze-one-quarter-of-the-city-2010-3>.
- Tabellini, G., 2010. Culture and institutions: economic development in the regions of Europe. *J. Eur. Econ. Assoc.* 8 (4), 677–716.
- Taylor, P., Fry, R., 2012. The rise of residential segregation by income: Pew Research Center Report, August 1.

- Topa, G., Zenou, Y., this volume. Neighborhood versus network effects. In: Duranton, G., Vernon Henderson, J., Strange, W.C. (Eds.), *Handbook of Regional and Urban Economics*, vol. 5. UN-Habitat, 2013. *State of the World's Cities 2012/2013: Prosperity of Cities*. Routledge Taylor & Francis Group, New York, 711 Third Avenue.
- Vigdor, J.L., 2010. Is urban decay bad? Is urban revitalization bad too? *J. Urban Econ.* 68 (3), 277–289.
- Vincent, T., 2005. Harlem. *Continuum Encyclopedia of Popular Music of the World: Locations*. N.p.: Continuum, Credo Reference. <http://search.credoreference.com/content/entry/contpmwl/harlem/0>.
- Wahl, F., 2013. Does medieval trade still matter? Historical trade centers, agglomeration and contemporary economic development. FZID Discussion Paper, No. 82-2013, November.
- Waldfoegel, J., 2003. Preference externalities: an empirical study of who benefits whom in differentiated-product markets. *RAND J. Econ.* 34 (3), 557–568.
- Waldfoegel, J., 2008. The median voter and the median consumer: local private goods and population composition. *J. Urban Econ.* 63 (2), 567–582.
- Waters, C.E.C., 1895. *Egypt*. Werner, New York.
- Watson, T., 2009. Inequality and the measurement of residential segregation by income in American neighborhoods. *Rev. Income Wealth* 55 (3), 820–844.
- Wheaton, W.C., 1977. Income and urban residence: an analysis of consumer demand for location. *Am. Econ. Rev.* 67, 620–631.
- Wheeler, C.H., La Jeunesse, E.A., 2007. Neighborhood income inequality. Federal Reserve Bank of St. Louis. Working paper 2006-039B, February.
- Williams, T., 2008. Harlem area is blighted, state agency declares. *The New York Times* (18 July). http://www.nytimes.com/2008/07/18/nyregion/18columbia.html?_r=0.
- Yoon, C., 2013. The decline of the rust belt: a dynamic spatial equilibrium analysis: Baruch College/CUNY Working paper, New York, 4 December.
- Zhang, J., 2004. A dynamic model of residential segregation. *J. Math. Sociol.* 28, 147–170.
- Zhang, J., 2011. Tipping and residential segregation: a unified schelling model. *J. Reg. Sci.* 51 (1), 167–193.

SECTION IV

Applied Urban Economics

This page intentionally left blank

CHAPTER 17

Taxes in Cities

Marius Brühlhart^{*,†}, Sam Bucovetsky[‡], Kurt Schmidheiny^{†,§,¶}

^{*}University of Lausanne, Lausanne, Switzerland

[†]Centre for Economic Policy Research (CEPR), London, UK

[‡]York University, Toronto, ON, Canada

[§]University of Basel, Basel, Switzerland

[¶]CESifo, Munich, Germany

Contents

17.1. Introduction	1124
17.2. Institutional Background	1126
17.2.1 Urban jurisdictional fragmentation across the world	1127
17.2.2 Local fiscal decentralization across the world	1132
17.2.3 Urban jurisdictional fragmentation and fiscal decentralization across the OECD	1139
17.2.4 Asymmetries in jurisdictional fragmentation	1140
17.2.5 Summary of institutional facts	1142
17.3. Tax Setting Across Asymmetric Jurisdictions	1145
17.3.1 Horizontal tax competition	1145
17.3.1.1 Theory	1145
17.3.1.2 Empirical studies	1150
17.3.2 Vertical tax competition	1152
17.3.3 Voting	1153
17.3.4 Central municipalities and suburbs	1156
17.4. Taxation and Urban Population Sorting	1161
17.4.1 Tax-induced urban population sorting: theory	1161
17.4.1.1 Aspatial general equilibrium models	1161
17.4.1.2 Asymmetry and centrality	1164
17.4.2 Tax-induced urban population sorting: Empirical studies	1166
17.4.2.1 Estimation of individual location choice	1167
17.4.2.2 Estimation of individual location choice in equilibrium	1168
17.4.2.3 Structural estimation	1170
17.5. Taxation and Agglomeration Economies	1171
17.5.1 Theory	1172
17.5.2 Empirical studies	1173
17.5.2.1 Do agglomeration economies make firms more or less sensitive to local taxation?	1174
17.5.2.2 Do local-level tax policies take account of agglomeration economies?	1176
17.5.2.3 Do agglomeration economies affect the equilibrium tax competition outcome significantly?	1178
17.6. Concluding Remarks	1178
Appendix	1179
Acknowledgments	1191
References	1191

Abstract

Most cities enjoy some autonomy over how they tax their residents, and that autonomy is typically exercised by multiple municipal governments within a given city. In this chapter, we document patterns of city-level taxation across countries, and we review the literature on a number of salient features affecting local tax setting in an urban context. In OECD countries, urban local governments on average raise some 10% of total tax revenue, and in non-OECD countries, they raise around half that share. We show that most cities are highly fragmented: urban areas with more than 500,000 inhabitants are divided into 74 local jurisdictions on average. The vast majority of these cities are characterized by a central municipality that strongly dominates the remaining jurisdictions in terms of population. These empirical regularities imply that analyses of urban taxation need to take account of three particular features: interdependence among tax-setting authorities (horizontally and vertically), jurisdictional size asymmetries, and the potential for agglomeration economies. We survey the relevant theoretical and empirical literatures, focusing in particular on models of asymmetric tax competition, of taxation and income sorting, and of taxation in the presence of agglomeration rents.

Keywords

Cities, Taxes, Tax competition, Fiscal federalism, Agglomeration, Sorting

JEL Classification Codes

H71, H73, R28, R51

17.1. INTRODUCTION

Cities the world over are big collectors of taxes. In the Organization for Economic Cooperation and Development (OECD), local governments raise about 13% of total tax revenue, and close to 80% of the population lives in cities. Hence, as a rough approximation, 1 in 10 tax dollars, on average, is raised by urban local governments. That share is lower in non-OECD countries, but with fiscal decentralization and urbanization both progressing in most developing countries, convergence toward OECD levels would seem to be only a matter of time.

While taxation by city governments is important on average, it is also very heterogeneous, even among developed countries. In the OECD, the local share of total tax revenue ranges from 1% to 33%. The set of tax instruments available to local governments differs as well, as does the degree of autonomy allowed to individual municipalities in setting their own tax rates.

The city-level tax authority typically represents the bottom of a federal pyramid of tax-raising government layers. Most countries in addition feature taxation by intermediate-level (state, provincial, cantonal, etc.) authorities as well as by the national government. Hence, city-level taxation will be characterized by interactions in at least three dimensions: among local governments within a city, among cities, and between local governments and the upper government layers.

In this chapter, we focus on characteristics that are of particular importance to city-level taxes and that can differ substantially from those of international fiscal relations.¹ We note three special features of city-level taxation, each of which we seek to address:

1. **Interdependence:** Given the comparatively small spatial scale of cities and the resulting mobility of the tax base, tax decisions by local governments are “horizontally” interdependent. Moreover, because all city-level governments coexist with one or several layers of governments above them, “vertical” interdependencies among government tiers must be considered as well. These interdependencies are furthermore affected almost everywhere by fiscal equalization schemes that redistribute tax revenue horizontally and vertically.
2. **Asymmetry:** Cities typically consist of a large central jurisdiction and several smaller surrounding jurisdictions. Central and noncentral jurisdictions differ in a number of ways, but a large disparity in economic and population size is of first-order relevance everywhere.
3. **Agglomeration:** One taxpayer’s location decision within and between cities is often linked to the location decisions of other agents. Firms seek proximity to each other in local clusters, and people often prefer to live near other people like them. Such agglomeration and sorting phenomena influence and are influenced by local tax policy.

A note on terminology: in this chapter, we use the term “city” in a geographic rather than in a political sense; that is, we use this term to denote a functional urban area, synonymous to terms such as “metropolitan area,” “urban local labor market,” or “travel-to-work area.” Cities contain multiple contiguous political jurisdictions. We refer to these jurisdictions as “municipalities,” and to the taxes they collect as “municipal” or “local” taxes. Among the municipalities within a city, we distinguish the “central municipality,” or simply “the center,” from its fringe of “noncentral” or “suburban” municipalities.²

What intellectual case is there for cities to raise their own tax revenues? The most frequently invoked analytical framework is the “decentralization theorem” formalized by [Oates \(1972\)](#) and [Besley and Coate \(2003\)](#). This approach highlights the trade-off between, on the one hand, spatially differentiating tax policy so as to satisfy heterogeneous voter preferences, and, on the other hand, internalizing fiscal spillover effects across jurisdictions. Considering in addition that taxpayers are mobile within and between cities, decentralized taxation can be efficient as taxpayers “vote with their feet” ([Tiebout, 1956](#)).

¹ [Wilson \(1999\)](#), [Gordon and Hines \(2002\)](#), [Brueckner \(2003\)](#), [Epple and Nechyba \(2004\)](#), [Fuest et al. \(2005\)](#), [Keen and Konrad \(2013\)](#), and several of the chapters in [Ahmad and Brosio \(2006\)](#) are among the relatively recent surveys on intergovernmental fiscal relations without a specifically urban focus.

² We mostly treat local governments as if they were of a single type, thereby abstracting from functional jurisdictions such as school or other special-purpose districts, whose boundaries might not overlap. In a seminal theoretical treatment, [Hochman et al. \(1995\)](#) show that differences in optimal spatial scopes across types of local public goods strengthen the case for city-level jurisdictional consolidation. For a discussion of the merits of multiple functional local jurisdictions, see, for example, [Frey and Eichenberger \(1996\)](#).

The interdependencies traditionally modeled in this literature are expenditure spillovers, but [Brueckner \(2004\)](#) shows that they could just as well be conceived of as interdependencies due to competition over a mobile tax base. Both these issues feature particularly prominently in the within-city context. As the case for local taxation is the stronger the larger are the spatial differences in voter preferences and the weaker are spillovers, much will depend on the within-city heterogeneity of voter preferences. This chapter will therefore pay particular attention to decentralized taxation acting as a trigger for spatial population sorting.

Decentralization of the tax authority from the central or regional level to local governments will furthermore have different implications depending on the size distribution of local jurisdictions. The decentralization theorem applies in this dimension too: more jurisdictionally fragmented cities will be able to cater better to local differences in preferences, at the cost of incomplete internalization of spillovers. In addition, greater jurisdictional fragmentation can imply efficiency losses in the provision of public goods that are subject to scale economies ([Alesina et al., 2004](#)), but it can also generate efficiency gains through Tiebout sorting and yardstick competition ([Hoxby, 2001](#)). In this chapter, we consider not only fragmentation itself but also the degree of asymmetry in jurisdictional size distributions and how that interacts with agglomeration economies.

We restrict this chapter to features we consider to be particularly germane to city-level taxation and which have been the subject of recent scholarly research. We therefore have no claim to a general treatment of urban public finance, but we can point readers to some excellent complementary surveys. The big trade-offs inherent in fiscal decentralization are summarized by [Epple and Nechyba \(2004\)](#), and work on political and institutional determinants of city-level policies is reviewed by [Helsley \(2004\)](#). For a survey focusing specifically on urban housing markets and tax capitalization, see [Ross and Yinger \(1999\)](#). Research on fiscal equalization policies, which constrain local fiscal autonomy in many countries, is reviewed in [Boadway \(2004\)](#). [Glaeser \(2013\)](#) offers a survey focusing on three central features of urban taxation in the United States: property taxation, intergovernmental transfers, and balanced-budget rules. Finally, the chapters by Gyourko and Molloy and Olsen and Zabel in this volume review research on the determinants and effects of city-level regulatory policies with respect to land and housing.

The chapter is organized as follows. We begin by providing a cross-country description of intracity allocations of the authority. In the remainder of the chapter, we survey the theoretical literature and empirical literature on the three topics we consider most relevant for city-level tax policy: jurisdictional asymmetry, population sorting, and agglomeration economies. The final section offers a concluding summary.

17.2. INSTITUTIONAL BACKGROUND

This section documents city-level fiscal decentralization patterns and thereby shows how tax competition among different jurisdictions *within* the same city potentially takes place in many countries around the world.

We see two prerequisites for tax competition to occur within a city: First, the city needs to be divided into several local jurisdictions—that is, municipalities. Second, the local jurisdictions need to have significant autonomy to raise local taxes. We study the first prerequisite in [Section 17.2.1](#) for 28 OECD countries, and we document the second prerequisite for 40 OECD and non-OECD countries in [Section 17.2.2](#). [Section 17.2.3](#) combines the results of the two previous sections and identifies countries with good conditions for intraurban tax competition. [Section 17.2.4](#) explores asymmetries in the size of local jurisdictions.

While our analysis is limited by the availability of informative data, we establish that decentralized taxation in cities is a ubiquitous phenomenon. The stylized facts from our analysis are summarized in [Section 17.2.5](#).

17.2.1 Urban jurisdictional fragmentation across the world

This subsection documents to what degree cities around the world are fragmented into local jurisdictions that potentially compete in tax levels. We think of cities as large *functional urban areas* that typically stretch across different administrative government units. A key challenge is therefore to work out an operational definition of cities that is consistent across countries. Given the available data, we shall consider only *large* cities, defined as functional urban areas with more than 500,000 inhabitants.

We use recently collected data from a collaborative venture by the OECD and the European Commission (EC).³ The OECD/EC definition of functional urban areas comprises three steps. Step 1 partitions sample country surfaces into 1 km² grid cells and identifies as high-density cells those with a population density greater than 1500 inhabitants per square kilometer on the basis of categorized satellite images. Step 2 generates clusters of contiguous (sharing at least one border) high-density cells. Low-density cells surrounded by high-density cells are added. Clusters with a total population of at least 50,000 inhabitants are identified as urban cores. Step 3 uses administrative data to calculate commuting flows from local administrative units (municipalities) into urban cores.⁴ Local administrative units with at least 15% of employed individuals in an urban core are assigned to the urban core. A contiguous set of assigned local administrative units forms a larger urban zone or functional urban area—that is, a “city” in our terminology. Non-contiguous urban cores with bilateral commuting flows of more than 15% of employed individuals are combined into a single polycentric larger urban zone.⁵ An example is provided in [Figure 17.1](#), where the single panels illustrate the three steps for the case of the Berlin area in Germany. The public OECD data contain information on all functional urban areas with total population above 500,000 inhabitants.

³ See [Brezzi et al. \(2012\)](#) and [Dijkstra and Poelman \(2012\)](#) for a detailed description of the data collection method.

⁴ In the United States, counties are used.

⁵ A threshold of 50% instead of 15% is applied as an exception for the United States. See ([OECD, 2013b](#)).

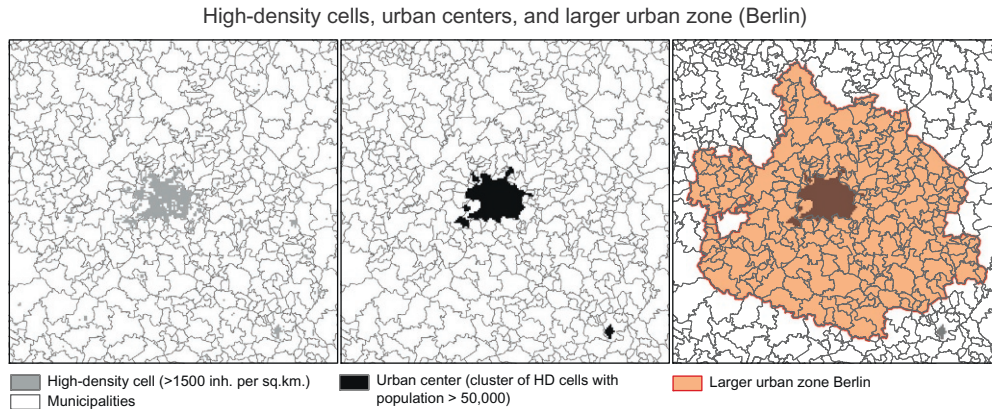


Figure 17.1 Construction of the Berlin functional urban area. The left panel shows the high-density cells with more than 1500 inhabitants per square kilometer and administrative municipal boundaries. The middle panel illustrates the construction of urban centers with a total population of more than 50,000 inhabitants. The right panel shows the construction of the larger urban zone based on bilateral commuting flows. *Source: European Commission, Directorate-General Regional and Urban Policy.*

This OECD/EC definition of urban areas has important advantages over the use of population data for administratively defined cities. Most importantly for our purpose, the definition is largely identical across countries. The procedure also identifies cities that straddle national borders, such as Geneva or Basel. Finally, the OECD/EC data are complementary to other approaches which draw on fine-grained satellite images of population clusters (Rozenfeld et al., 2008, 2011) or night-lights (Henderson et al., 2012) to define “cities,” but which do not include economic linkages across space such as commuting flows. The OECD/EC definition therefore represents a uniquely suitable operationalization of functional urban areas for our purpose.⁶

The OECD defines a “local government” as the lowest level of general-purpose government with relevant responsibilities.⁷ It explicitly does not consider special purpose jurisdictions such as school districts. See Table A.1 for a description of the local

⁶ For a critical discussion of city definitions and an alternative algorithm based solely on commuting patterns, see Duranton (2013).

⁷ The exact OECD criterion for “local governments” is as follows: “Have only one level of local government per country, notably the lowest tier (even if more than one level of government may have relevant responsibilities over the same territory). Identify only general-purpose local governments, excluding the specific function governments (for example, school district, health agencies, etc.). United Kingdom: For those areas where the County Councils were abolished the local authority (either a Metropolitan District Council or a Unitary District Council) is used. For London, the Borough Councils are used. United States: In the geographic areas where municipalities or townships do not represent a general purpose government, the county governments were considered” (OECD, 2013a, p. 174).

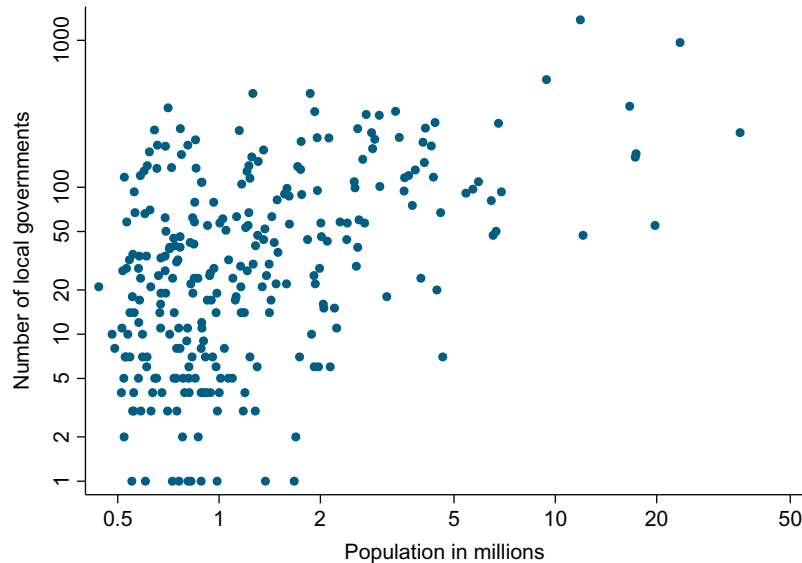


Figure 17.2 Number of municipalities and population across 275 cities in 28 OECD countries, 2012.
Source: OECD (Regional Statistics).

government level in each country. We shall use the term “municipality” to refer to all of these country-specific types of local jurisdiction.

Figure 17.2 plots the number of municipalities against the population of all sampled functional urban areas in 28 OECD countries. The first observation is that only 10 cities (all of them in Mexico) are entirely contained in one municipality. The other 265 urban areas are fragmented into up to 1375 municipalities (Paris, France). Table 17.1 shows the cities with the highest number of municipalities. The top 10 cities are found in France (Paris and other cities), Korea (Seoul), the United States (Chicago and other cities), Austria (Vienna), and the Czech Republic (Prague). The most fragmented cities in the other sampled OECD countries are also listed in Table 17.1. The average OECD city is divided into 74 municipalities.

Not surprisingly, larger cities tend to contain more municipalities. A regression of the logarithm of the number of municipalities on the logarithm of population yields a highly significant slope coefficient of 0.90 (standard error 0.10, $p < 0.001$). The OECD therefore also reports an index of urban fragmentation which takes account of the different sizes of cities: the number of municipalities per 100,000 inhabitants in the city. Table 17.1 also shows the top urban areas with respect to this index. We again find a number of French cities among the top 10 (e.g., Rouen with 49 local governments per 100,000 inhabitants), but also cities from the Czech Republic (Brno), Austria (Graz), Spain (Saragossa), and Switzerland (Geneva).

Table 17.1 Most fragmented OECD cities
Number of local jurisdictions

Number of local jurisdictions				Local governments per 100,000			
Rank	Urban area	ISO country code	Number	Rank	Urban area	ISO country code	Index
1	Paris	FRA	1375	1	Rouen	FRA	49.06
2	Seoul Incheon	KOR	965	2	Brno	CZE	38.13
3	Chicago	USA	540	3	Toulouse	FRA	34.39
4	Prague	CZE	435	4	Strasbourg	FRA	32.57
5	Toulouse	FRA	434	5	Grenoble	FRA	29.42
6	New York	USA	356	6	Graz	AUT	28.04
7	Rouen	FRA	346	7	Rennes	FRA	27.45
8	Minneapolis	USA	329	8	Saragossa	ESP	24.67
9	Lyon	FRA	327	9	Geneva	CHE	23.9
10	Vienna	AUT	313	10	Prague	CZE	23.28
...				...			
11	Hamburg	DEU	308	14	Wichita	USA	21.54
13	Madrid	ESP	272	18	Bratislava	SVK	18.83
14	Milan	ITA	252	29	Porto	PRT	11.48
19	Lisbon	PRT	235	34	Augsburg	DEU	10.96
19	Tokyo	JPN	235	52	Quebec	CAN	6.87
28	Geneva	CHE	193	53	Budapest	HUN	6.39
30	Montreal	CAN	191	54	Busan	KOR	6.35
32	Budapest	HUN	183	58	Liege	BEL	6.14
45	Bratislava	SVK	136	59	Milan	ITA	6.13
63	Warsaw	POL	101	60	Oaxaca de Juárez	MEX	6.01
64	Brussels	BEL	99	77	Tallinn	EST	5.28
68	Athens	GRC	94	83	Lublin	POL	4.92
95	Amsterdam	NLD	57	86	Ljubljana	SVN	4.86
95	Copenhagen	DNK	57	118	Thessaloníki	GRC	2.9
100	Mexico City	MEX	55	121	Copenhagen	DNK	2.84
108	Santiago	CHL	47	125	Eindhoven	NLD	2.74
108	London	GBR	47	132	Tokushima	JPN	2.5

Top 10 or top of country in respective ranking.

Source: OECD (Regional Statistics), various years.

We next turn to a comparison of urban jurisdictional fragmentation across countries. [Figure 17.3](#) shows the average number of municipalities per city for each country. The first and most important observation is that urban areas are substantially fragmented in all OECD countries. Even in the lowest-ranked country (Ireland), the only sampled city (Dublin) is fragmented into seven local governments. France tops the OECD countries, with on average 280 municipalities per city, followed by the Czech Republic, Austria, and Portugal. The right panel in [Figure 17.3](#) shows that there is substantial variation

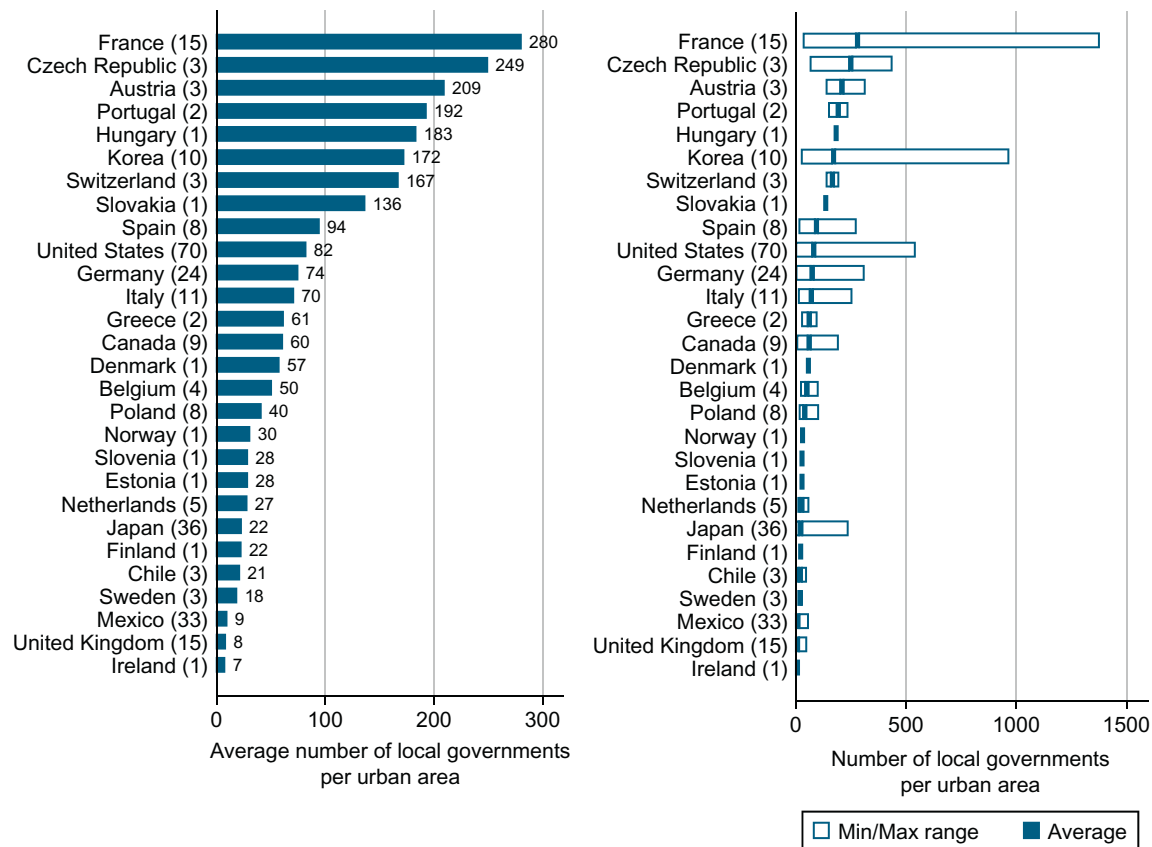


Figure 17.3 Number of municipalities within OECD cities. Number of sample cities in parentheses. *Source: Own calculations based on OECD (Regional Statistics).*

within many countries. In France the range is from 35 to 1375, in Korea it is from 27 to 965, and in the United States it is from 2 to 540. The variation within counties is larger than that across counties (within-country standard deviation 109, between-country standard deviation 80).

Figure 17.4 shows country averages of the number of municipalities per 100,000 inhabitants.⁸ The Czech Republic appears as the country with the highest average degree of jurisdictional fragmentation, with 24 local governments per 100,000 inhabitants. It is followed by France (21), Austria (21), and Switzerland (19). There is again substantial variation *within* countries, although now smaller than *between* countries (within-country standard deviation 4.2, between-country standard deviation 7.2). For example, there are between six (Toulon) and 49 (Rennes) municipalities per 100,000 inhabitants in French urban areas, between two (Barcelona) and 25 (Saragossa) municipalities per 100,000 inhabitants in Spain and between 0.2 (Tampa, Florida) and 22 (Wichita, Kansas) municipalities per 100,000 inhabitants in the United States.

In sum, almost all cities in OECD countries with a population above 500,000 inhabitants are fragmented into several local governments (municipalities). On average, there are 74 municipalities per city and 4.9 municipalities per 100,000 inhabitants. We explore in the next section to what extent these local governments can autonomously set local taxes.

17.2.2 Local fiscal decentralization across the world

This subsection documents the degree of local fiscal decentralization in 40 countries across the world. Our principal data source are the Government Finance Statistics (GFS) collected by the International Monetary Fund (IMF).⁹ The GFS report revenue and expenditure data of different government units for countries around the world. For many countries, the data are reported separately for different levels of government: the central government, state governments and local governments. In addition, aggregates for *general government* are reported, eliminating double counting from transfers between government layers. This data source has been widely used to document the degree of government decentralization (e.g., Arzaghi and Henderson, 2005; Stegarescu, 2006) at the subcentral (state plus local) level. We will focus on the *local* (municipal) level, which corresponds most closely to the definition we use to document urban jurisdictional fragmentation in Section 17.2.1. The GFS do not report spatially

⁸ We use unweighted averages. OECD (2013a, p. 47) reports averages weighted by the population of the urban area. The unweighted average is more informative to document how many cities are fragmented to what degree.

⁹ See International Monetary Fund (2001) for a detailed description of the data-collecting process. The OECD Fiscal Decentralization Database is an alternative data source with very similar information about OECD countries. We use the IMF GFS because they cover more countries and are more transparently documented.

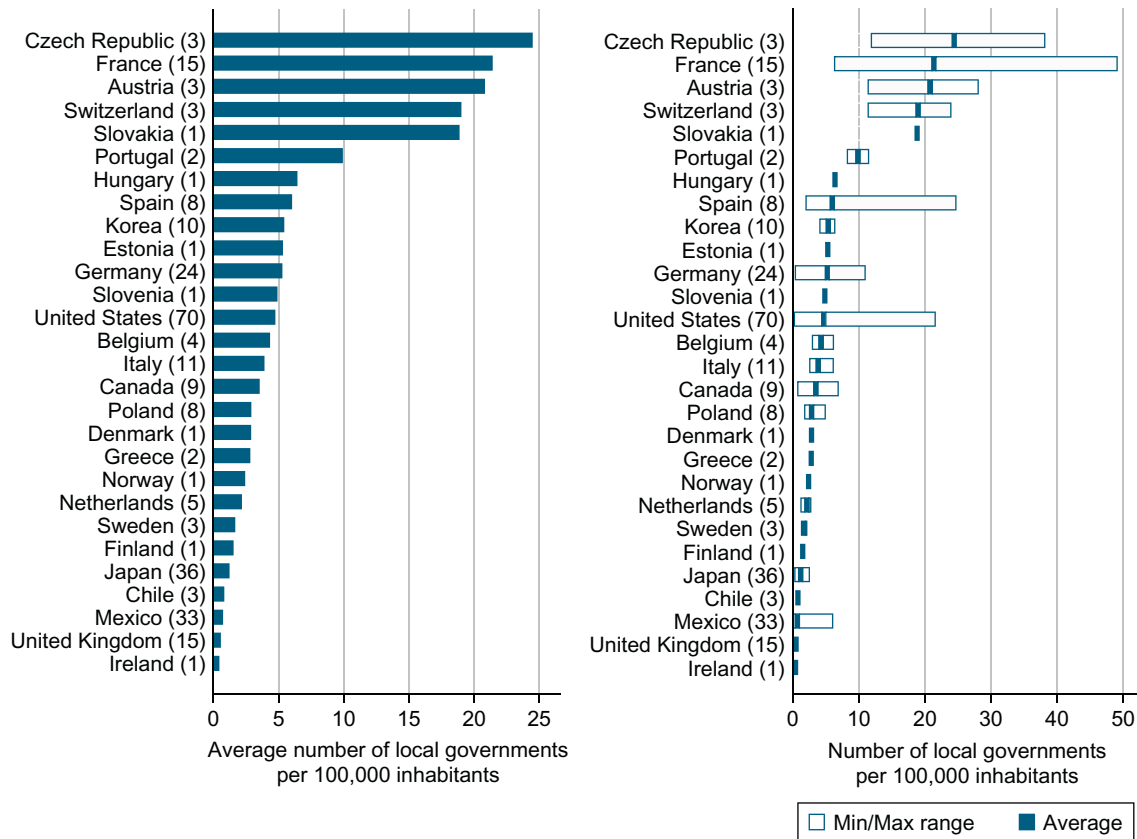


Figure 17.4 Jurisdictional fragmentation of OECD cities. Number of sample cities in parentheses. *Source: Own calculations based on data from OECD (Regional Statistics).*

disaggregated data for the United States for recent years. We therefore use the historical database on individual government finances (IndFin) from the US Census Bureau and replicate the GFS definitions of revenue and expenditure categories as closely as possible.¹⁰

We can identify municipal tax revenue for 40 countries. Using the World Bank country classification, we observe data from 17 high-income OECD countries, 4 other high-income countries, 12 upper-middle-income countries, and 7 lower-middle-income countries.¹¹ Unfortunately, there are, in addition, a number of countries for which the GFS do not distinguish the regional (state) and local (municipal) levels.¹² For example, “local” data for France include 26 regions and 100 departments in addition to 36,000 communes, and “local” data for Sweden include 20 county councils in addition to 116 municipal associations and 290 municipalities. Table A.3 lists the local government units which are used in our cross-country sample. For each country, we use the most recent year for which revenue information is available at the local level. The observed years range from 2012 (United Kingdom) to 2003 (Swaziland). Table A.3 reports the observed years for all sample countries.

Stegarescu (2006) proposes several indices of fiscal decentralization at the subcentral (state and municipal) level. We apply these indices to the local (municipal) level. The first index is local government tax revenue as a share of general government tax revenue:

$$\text{LTS} = \frac{\text{local government tax revenue}}{\text{general government tax revenue}}, \quad (17.1)$$

where LTS stands for local tax share.

The index LTS is calculated from the GFS/IndFin data. A detailed description of the variables used is provided in the Appendix. Figure 17.5 shows the LTS for the 40 countries in our sample; exact numbers and averages are reported in Table A.5. Finland ranks top, with 33% of tax revenue collected at the local level. Iceland, Estonia, Switzerland, Latvia,

¹⁰ For the years 1987 and 1992, we have local data for the United States from both GFS and IndFin. The decentralization indices based on IndFin data are systematically smaller than those based on GFS data. However, the differences do not substantially alter the ranking of tax decentralization across countries. See the Appendix for a detailed comparison.

¹¹ We use the World Bank list of economies (February 2014) available online at <http://siteresources.worldbank.org/DATASTATISTICS/Resources/CLASS.XLS>. The World Bank divides economies according to 2012 gross national income *per capita* into four income groups: low income, USD 1035 or less; lower-middle income, USD 1036 to USD 4085; upper-middle income, USD 4086 to USD 12,615; and high income, USD 12,616 or more.

¹² The excluded countries are Afghanistan, Azerbaijan, Belarus, China, Republic of the Congo, Costa Rica, Croatia, Cyprus, Czech Republic, Denmark, France, Georgia, Hungary, Ireland, Israel, Italy, Japan, Kazakhstan, Korea, Lesotho, Mauritius, Mexico, Moldova, Mongolia, Morocco, Netherlands, Norway, Poland, Portugal, Romania, Serbia, Slovakia, Sweden, Tajikistan, Thailand, Tunisia, Turkey, Ukraine, and the West Bank and Gaza. See Table A.4 for a description of the problematic regional units.

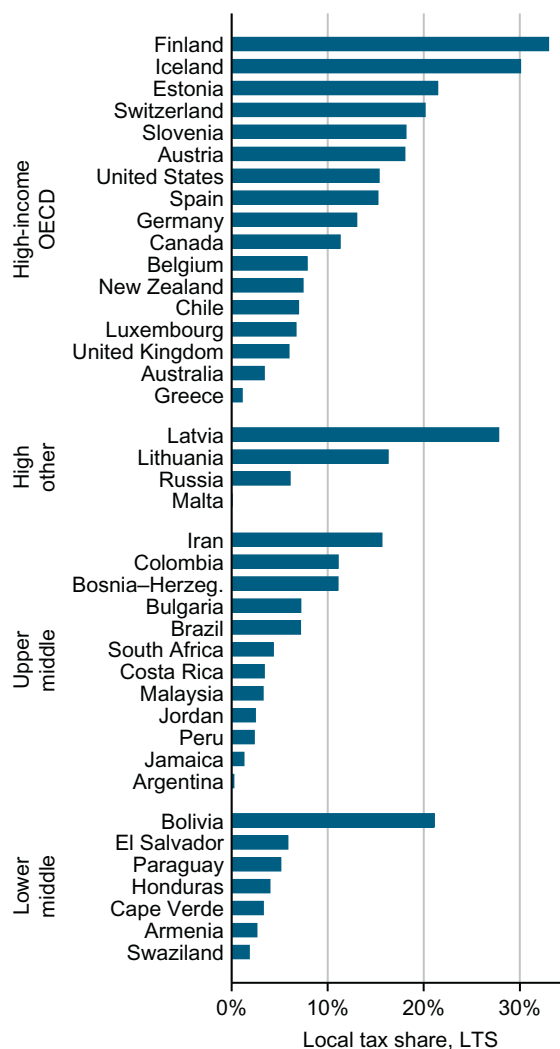


Figure 17.5 Decentralization of tax revenue at local (municipal) level. Local tax revenue as share of general government tax revenue. *Source: Own calculations based on data from the IMF (GFS) and US Census (IndFin).*

and Bolivia also have local tax revenue shares of over 20%. At the bottom of this list, we find Swaziland, Jamaica, Greece, Argentina, and Malta, with local tax revenue shares below 2%. On average, 10% of tax revenue is collected at the local level in our sample (13% in OECD countries). [Figure 17.5](#) clearly shows that many countries other than the United States feature substantial tax collection at the local level.

Tax competition can take place at the local level only if those jurisdictions have real autonomy over the revenue they collect. The [OECD \(1999\)](#) therefore classifies local taxes

into nine groups with decreasing local autonomy over determining the tax rate and tax base. Table 17.2 shows the nine categories relabeled for local governments instead of all subcentral government levels. According to the classification in Table 17.2, only tax groups (a)–(c) are relevant local tax revenue. Stegarescu (2006) proposes calculating local governments' tax revenue in groups (a)–(c) as a fraction of general government tax revenue:

$$\text{ALTS} = \frac{\text{local government tax revenue (a) to (c)}}{\text{general government tax revenue}}, \quad (17.2)$$

where ALTS stands for autonomous LTS.

The GFS/IndFin data do not report the degree of autonomy in tax setting. We therefore draw on additional data sources. On the basis of surveys of national financial laws and constitutions, Blöchliger and Rabesona (2009) report the share of local tax revenue in each of tax groups (a)–(e) in Table 17.2. For the United States, we use similar data from Stegarescu (2006) because Blöchliger and Rabesona (2009) do not distinguish local tax groups for the United States. These data are available for 14 high-income countries out of the 40 countries in our initial sample. We can therefore calculate ALTS as

$$\text{ALTS} = \frac{\text{local government tax revenue} \times \text{share in classes (a)–(c)}}{\text{general government tax revenue}}. \quad (17.3)$$

Table A.6 presents the reported shares of the individual tax categories.

Figure 17.6 shows ALTS for 15 high-income countries; the exact numbers are reported in Table A.5. For most countries, ALTS is very similar to LTS. In these countries, most of the local tax revenue belongs to tax classes (a)–(c). For two countries, however, the effective local tax autonomy is dramatically lower than that reported in LTS: autonomous local tax revenue in Austria is only 1.5% of general government tax revenue rather than the 18% when counting all local tax revenue, and in New Zealand ALTS is 0% instead of 7.4%. Autonomous local tax revenue is also somewhat smaller in Spain and Germany. Finland leads the ranking in autonomous tax decentralization. There are still

Table 17.2 Classification of local (municipal) taxes

(a)	LG determines tax rate and tax base
(b)	LG determines tax rate only
(c)	LG determines tax base only
(d)	Tax sharing
(d.1)	LG determines revenue split
(d.2)	Revenue split changed only with consent of LG
(d.3)	Revenue split changed unilaterally by CRG (legislation)
(d.4)	Revenue split changed unilaterally by CRG (annual budget)
(e)	CRG determine tax rate and tax base

LG, local (municipal) government; CRG, central and/or regional governments.

Source: Adapted from (OECD, 1999, p. 11).

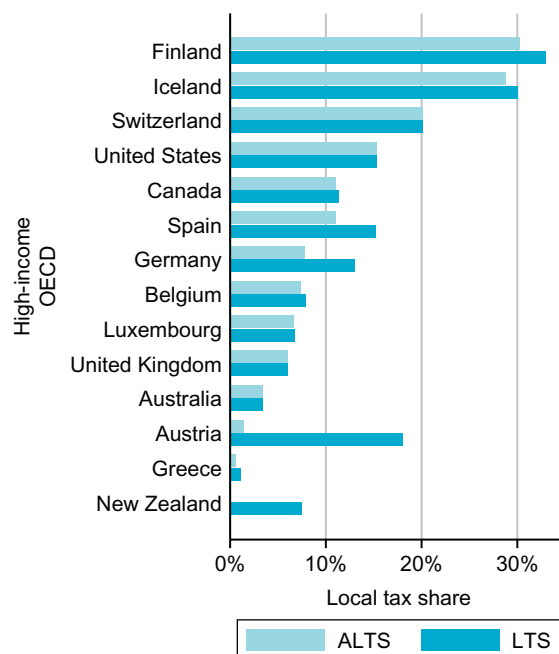


Figure 17.6 Decentralization of tax revenue at the local (municipal) level. Local tax revenue with real tax-raising autonomy as a share of general government tax revenue (ALTS). *Source: Own calculations based on data from the IMF (GFS), US Census (IndFin), and OECD.*

three countries (Finland, Iceland, and Switzerland) with local tax revenue shares above 20% of global tax revenue and six (additionally the United States, Canada, and Spain) with shares above 10%.

The GFS/IndFin data furthermore allow us to decompose local tax revenue into different tax sources: tax on household income, corporate income, property, and consumption of goods and services (including sales and value-added tax) and other tax bases. [Figure 17.7](#) illustrates the composition of the total LTS for 38 countries; the exact numbers are reported in [Table A.5](#). In our sample of countries, *property taxes* are the most important source of local revenue, with an average share of 43%, followed by personal income taxes (21%), consumption taxes (21%), other taxes (8%), and corporate income taxes (5%). Among the 16 countries with the highest degree of fiscal decentralization (LTS > 10%), however, personal income taxes dominate, with an average share of 42%, followed by property taxes (25%), consumption taxes (21%), other taxes (7%), and corporate income taxes (5%).

History appears to play a significant role in explaining which tax instruments are used by local governments. Property taxation is most important in the former British Empire. In our five “Anglo-Saxon” sample countries (Australia, Canada, the United Kingdom, the United States, and New Zealand), the local revenue share of property taxation ranges

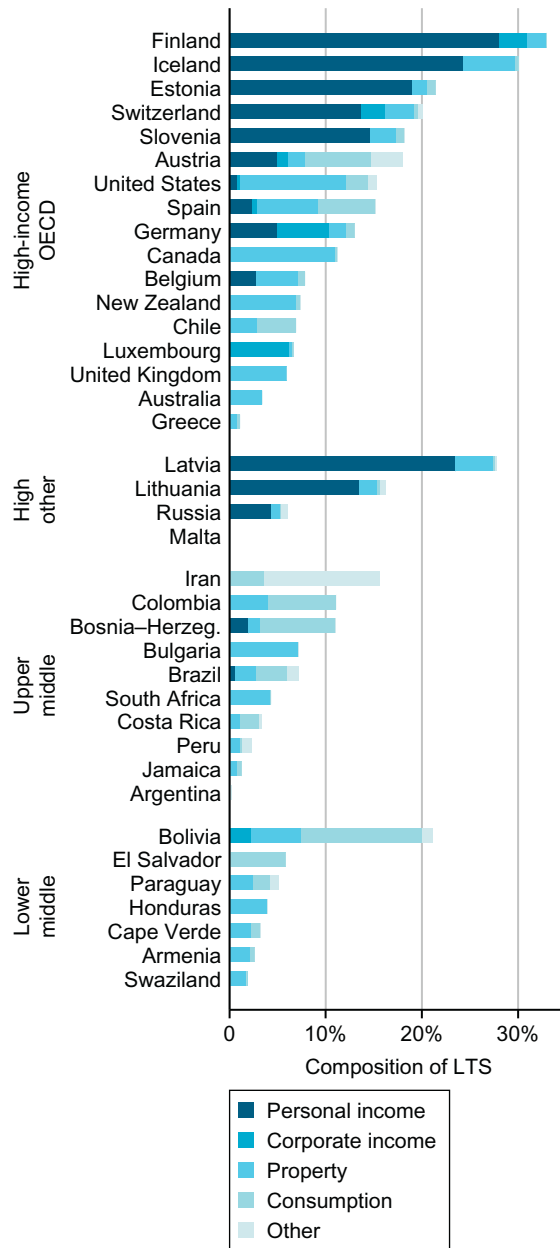


Figure 17.7 Decomposition of local (municipal) tax revenue into tax on personal income, corporate income, property, consumption, and other tax bases. Consumption taxes are taxes on goods and services, including sales, motor vehicle, and alcohol taxes. Local tax revenue is reported as a share of total general government tax revenue. *Source: Own calculations based on data from the IMF (GFS) and US Census (IndFin).*

from 72% to 100%. This share exceeds 50% in only two of the remaining 12 sample OECD countries (Belgium, 55%, and Greece, 75%).

17.2.3 Urban jurisdictional fragmentation and fiscal decentralization across the OECD

Section 17.2.1 documented how cities in OECD countries are fragmented into a multitude of local governments. In Section 17.2.2, we showed that local governments have substantial taxing powers in many countries around the world. Here, we combine the information of these two sections.

We have data on both local jurisdictional fragmentation and fiscal decentralization for 13 OECD countries. For 10 of them we observe the more informative decentralization index ALTS, and for 3 only the index LTS. Figure 17.2 shows the position of all 13 countries in the space of fragmentation and decentralization. Tax competition *within* cities can take place only if cities are fragmented into municipalities with some autonomy over local tax rates and/or the tax base. In Figure 17.8, this is the northeast corner, where Switzerland is positioned. Austria has a higher degree of jurisdictional fragmentation than Switzerland, but Austrian *Gemeinden* have no real local taxing power. Finland has a higher degree of tax decentralization but Finnish cities are fragmented into relatively few *Kommuner*. We also identify a group of countries

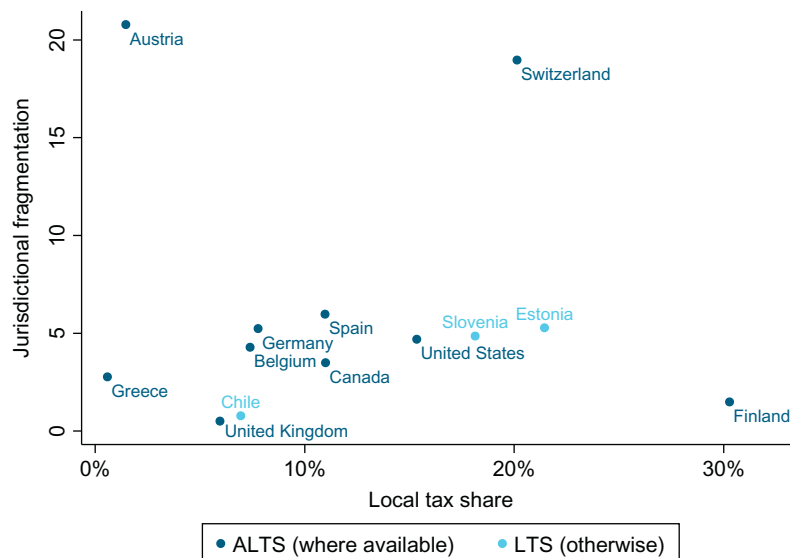


Figure 17.8 Jurisdictional fragmentation (average number of municipalities per 100,000 inhabitants, 2012) and local tax decentralization (local tax revenue as a share of general government tax revenue, various years). ALTS considers local tax revenue with real tax autonomy; LTS considers all local tax revenue. Source: Own calculations based on data from the OECD (Regional Statistics), IMF (GFS), and US Census (IndFin).

(the United States, Spain, Canada, Germany, and Belgium) with substantial values in both fragmentation and decentralization. Slovenia and Estonia also belong to this group, although we do not know the assignment of real taxing autonomy in these countries. Chile, the United Kingdom, and Greece have a low degree of both fragmentation and decentralization. However, even in this group of countries there may very well be scope for local-level tax competition.

Note that the sample of countries in [Figure 17.8](#) is determined by the availability of comparative cross-country data. There are likely many more countries with good conditions for urban tax competition. In particular, we cannot include many OECD countries with a high degree of subcentral tax decentralization but where the municipal share is unknown, such as Sweden, Denmark, Norway, Japan, France, and Italy. We also do not include all non-OECD countries with a substantial degree of fiscal decentralization, such as Latvia, Lithuania, Iran, Colombia, Bosnia–Herzegovina, and Bolivia (see [Section 17.2.2](#)).

17.2.4 Asymmetries in jurisdictional fragmentation

In this section, we study the size distribution of local jurisdictions *within* cities. Anecdotal evidence suggests that the core municipality in large urban areas typically “dominates” the urban area in terms of population. This section seeks to quantify this casual observation.

The OECD data do not report population figures for individual municipalities within cities. We therefore use alternative data provided by the EC for the year 2006 and the US Census for 2012 (see the Appendix for details). The EC data are based on the joint OECD/EC definition of cities, but unfortunately differ substantially in some details.¹³ The US data for 2012 complement the EC data with information on US cities.

We first calculate the population of the largest local jurisdiction as a share of the total population of the city. [Table 17.3](#) shows the European and US cities with the highest population share of the dominant municipality. The list is topped by the city of Saragossa in Spain, the main municipality of which hosts 93% of the city’s population. It is followed by Genoa in Italy (85%) and El Paso, Texas, in the United States (81%). On average, in our sample of 158 cities across 17 countries the population share of the largest municipality is 38.9%.

¹³ The EC data combine the German cities of Essen, Dortmund, Duisburg, and Bochum into the combined city “Ruhrgebiet,” which makes it the largest city in Germany. The number of municipalities (local governments) is identical or very similar in the EC and the OECD data, except for Saragossa (Spain), for which the OECD reports 210 municipalities and the EC reports 21, while both report a similar total population. All cities in the Czech Republic contain a significantly greater number of municipalities in the EC data than in the OECD data—for example, 435 and 729, respectively, for Prague. Because of this resulting lack of comparability with [Section 17.2.1](#), we do not include the Czech Republic in [Section 17.2.4](#).

Table 17.3 Population shares of largest municipalities

Rank	Urban area	ISO country code	Population	Local governments	Share of largest (%)
1	Saragossa	ESP	702,349	21	93.2
2	Genoa	ITA	736,058	38	84.5
3	El Paso	USA	830,827	7	81.1
4	Lodz	POL	967,581	17	78.5
5	Malaga	ESP	729,280	9	76.8
6	Wroclaw	POL	829,453	19	76.6
7	Tallinn	EST	536,059	24	73.3
8	Jacksonville	USA	1,190,394	14	70.3
9	Palermo	ITA	968,197	26	68.6
10	Berlin	DEU	4,980,394	246	67.5
...					
16	Vienna	AUT	2,599,439	313	63.6
18	Budapest	HUN	2,781,514	186	60.9
22	The Hague	NLD	796,581	7	59.0
30	Gothenburg	SWE	894,311	14	54.7
34	Ljubljana	SVN	485,374	26	52.1
41	Marseille	FRA	1,692,351	132	50.2
52	Oslo	NOR	1,113,227	34	47.3
56	Antwerp	BEL	1,014,444	32	45.6
60	Helsinki	FIN	1,248,302	14	43.2
102	Thessaloníki	GRC	996,428	29	31.2
103	Zurich	CHE	1,097,224	130	31.2
	Average		1,940,193	97.8	38.9

Top 10 or top of country in respective ranking.

Source: European data for 2006 from the EC (Urban Audit); US data for 2012 from the OECD (Regional Statistics) and US Census (Population Estimates Program).

The population share of the largest municipality needs to be compared with the share of the city's other municipalities. We calculate three measures to make such a comparison: the first measure is the ratio of the population of the largest municipality to the average population of all the city's municipalities. This measure would be 1 if all municipalities were of identical size. The second measure is the ratio of the population of the largest municipality to the population of the second largest municipality. This measure would also be 1 with identical municipality sizes, and it would be 2 if municipality sizes followed Zipf's rank-size rule.¹⁴ The third measure is the ratio of the population of the largest municipality to its theoretical population size if all the city's municipalities

¹⁴ Zipf (1949) postulated in his nonstochastic version that city sizes follow the rank-size rule:

$$\text{pop}_r = \text{pop}_1 / r,$$

where pop_r is the population of a city with rank r and pop_1 is the population of the largest city.

followed Zipf's rank-size rule.¹⁵ The third measure would be 1 if all municipality sizes followed the rank-size rule.

Table 17.4 shows the top-ranked cities according to each of the three measures. We see for all three measures that the top-ranked cities deviate starkly from both the uniform distribution rule and the rank-size rule. For example, the central municipality of Paris is 268 times bigger than the average municipality in the Paris urban area, the central municipality of Saragossa is 102 times bigger than the town of Zuera, which is the second largest municipality in the Saragossa urban area, and the central municipality of Berlin is 4.1 times bigger than what the rank-size rule would predict for the Berlin urban area.

Substantial deviations from both the uniform distribution rule and the rank-size rule can be found in all of the 17 sample countries. Figure 17.9 shows the average of the three measures for each country. The exact numbers along with minima and maxima are reported in Table A.7. At the top of the left panel is Hungary, where the largest municipality in the only urban area (Budapest) is 113 times larger than the average municipality; at the bottom is Finland, where the largest municipality in the only urban area (Helsinki) is 6 times larger than the average municipality. At the top of the right panel is Estonia, where the largest municipality in the only urban area (Tallinn) is 29 times bigger than the second largest municipality; at the bottom is Finland, where the largest municipality in Helsinki is only 2.4 times bigger than the second largest municipality.

17.2.5 Summary of institutional facts

Our analysis of data from the IMF (GFS), the OECD (Regional Statistics), Eurostat (Urban Audit), and the US Census (IndFin, Population Estimates Program) can be summarized in the following four stylized facts:

Result 17.2.1. *With the exception of some Mexican cities, all OECD cities with more than 500,000 inhabitants are fragmented into multiple local governments. On average, there are 74 local governments per functional urban area. The degree of urban jurisdictional fragmentation differs substantially both within and across countries.*

Result 17.2.2. *All of our 40 sample countries collect some tax revenue at the local (municipal) level. On average, 10.0% of the countries' total tax revenue is collected locally; 6 countries collect more than 20% locally, and 16 countries collect more than 10% locally. Considering*

¹⁵ The theoretical size of the largest municipality in an urban area with a given total population pop and given number of municipalities N is calculated as follows:

$$\text{pop}_1 = \text{pop} / [\psi(N+1) - \psi(1)],$$

where $\psi(\cdot)$ is the digamma function. $\psi(N+1) - \psi(1)$ equals the finite harmonic series $1 + 1/2 + \dots + 1/N$.

Table 17.4 Top asymmetry measures of urban areas

Largest vs. average-sized municipality				Largest vs. second largest jurisdiction				Largest vs. Zipf prediction			
Rank	Urban area	ISO country code	Ratio	Rank	Urban area	ISO country code	Ratio	Rank	Urban area	ISO country code	Ratio
1	Paris	FRA	268.1	1	Saragossa	ESP	102.1	1	Berlin	DEU	4.1
2	Hamburg	DEU	229.1	2	Genoa	ITA	57.2	2	Vienna	AUT	4.0
3	Vienna	AUT	199.0	3	Vienna	AUT	43.3	3	Rome	ITA	3.7
4	New York	USA	178.7	4	Baltimore	USA	41.4	4	Hamburg	DEU	3.6
5	Berlin	DEU	166.2	5	Jacksonville	USA	38.5	5	Genoa	ITA	3.6
6	Toulouse	FRA	163.7	6	Rome	ITA	33.1	6	Budapest	HUN	3.5
7	Chicago	USA	155.7	7	New York	USA	30.1	7	Wichita	USA	3.5
8	Budapest	HUN	113.2	8	Munich	DEU	29.3	8	Saragossa	ESP	3.4
9	Louisville	USA	96.5	9	Tallinn	EST	29.0	9	Louisville	USA	3.4
10	Rome	ITA	96.2	10	Wroclaw	POL	29.0	10	New York	USA	3.2
...				...				...			
13	Madrid	ESP	93.1	11	Budapest	HUN	28.1	12	Warsaw	POL	2.9
27	Warsaw	POL	52.1	19	Paris	FRA	19.1	15	Tallinn	EST	2.8
34	Geneva	CHE	43.8	39	Zurich	CHE	11.3	16	Marseille	FRA	2.7
72	Athens	GRC	20.8	43	Antwerp	BEL	10.4	52	Ljubljana	SVN	2.0
79	Tallinn	EST	17.6	52	Stockholm	SWE	8.4	58	Oslo	NOR	1.9
82	Oslo	NOR	16.1	56	Ljubljana	SVN	8.2	67	Antwerp	BEL	1.8
85	Antwerp	BEL	14.6	59	Rotterdam	NLD	8.1	70	Gothenburg	SWE	1.8
88	Amsterdam	NLD	14.1	92	Oslo	NOR	5.0	74	Rotterdam	NLD	1.7
91	Ljubljana	SVN	13.5	103	Athens	GRC	4.3	78	Zurich	CHE	1.7
104	Stockholm	SWE	10.5	135	Helsinki	FIN	2.4	97	Helsinki	FIN	1.4
136	Helsinki	FIN	6.1					115	Thessaloníki	GRC	1.2
	Average		33.7		Average		9.7		Average		1.7

Top 10 and top of country in respective ranking.

Source: European data for 2006 are from the EC (Urban Audit); US data for 2012 are from the OECD (Regional Statistics) and US Census (Population Estimates Program).

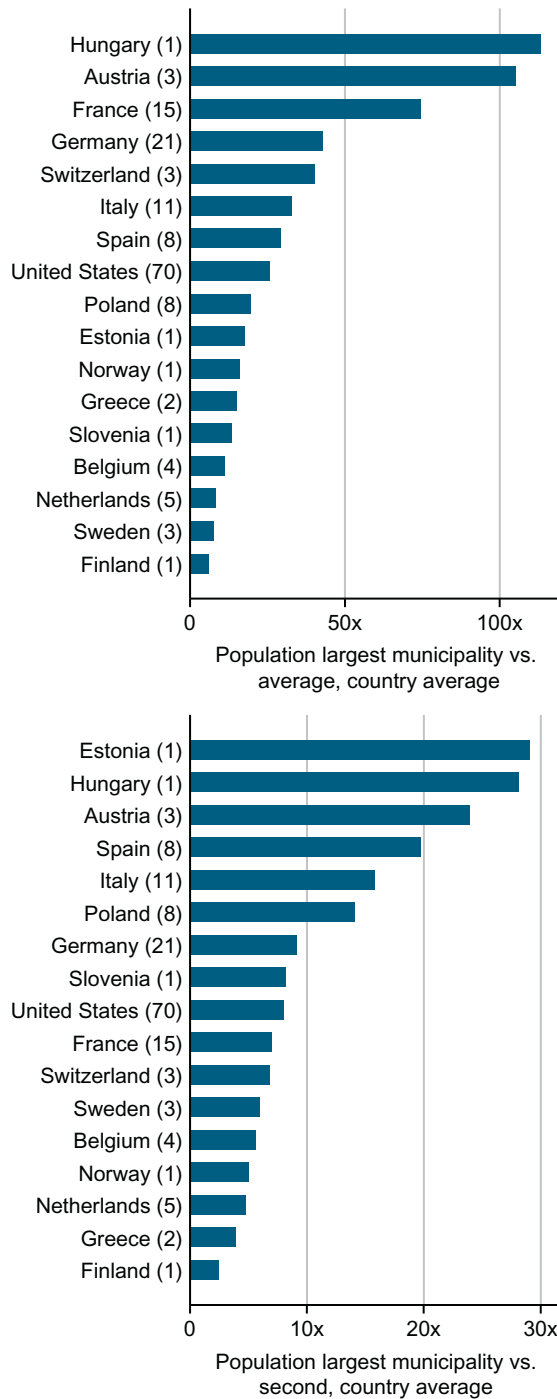


Figure 17.9 Asymmetries across local governments within cities. Number of sample cities in parentheses. *Source: Own calculations based on European data for 2006 from the EC (Urban Audit) and US data for 2012 from the OECD (Regional Statistics) and US Census (Population Estimates Program).*

local tax autonomy substantially reduces the effective degree of tax decentralization for some countries. The degree of local fiscal decentralization differs substantially both within and across countries.

Result 17.2.3. *We identify eight countries with conditions for intraurban tax competition at least as strong as those in the United States.*

Result 17.2.4. *Most OECD cities are characterized by a central municipality that strongly dominates the city in terms of population, beyond what would be predicted by Zipf's law.*

17.3. TAX SETTING ACROSS ASYMMETRIC JURISDICTIONS

Among the most striking stylized facts emphasized in [Section 17.2](#) are differences in the population share of the central municipality of a city, and in the number of municipalities in a city. Just as the structure of an industry affects the prices charged by firms, the structure of a city should influence the tax rates set by its municipalities.

17.3.1 Horizontal tax competition

17.3.1.1 Theory

If there were only one level of government, how should the size distribution of jurisdictions affect tax rates? Much of the theoretical (and empirical) analysis of this issue has used what will be referred to here as the “basic tax competition model,” developed in [Wilson \(1986\)](#), [Zodrow \(1986\)](#), and [Wildasin \(1988\)](#). In that model, capital is perfectly mobile within a city, and all other factors of production are immobile. This immobility assumption makes the basic tax competition model better suited to competition among larger units than to competition among municipalities within a single city. In the basic model, local public output benefits only the immobile residents, not owners of mobile capital. Allowing (some categories of) public expenditure to increase the return to capital will reduce or reverse governments’ incentives to attract capital by cutting tax rates. An even more important extension for urban public finance is to incorporate mobility of residents. Some attempts to do so are discussed in this subsection. However, to date there are no coherent and plausible models of taxation which incorporate these features. New approaches are needed to analyze more “urban” fiscal competition, approaches in which different levels of government, population sorting, and locational features play a greater role.

The basic tax competition model makes a strong prediction about how tax rates vary across the municipalities within a city: smaller municipalities have lower tax rates. [Bucovetsky \(1991\)](#) and [Wilson \(1991\)](#) provide a derivation of this result in the case of two jurisdictions, and [Wilson \(1999\)](#) offers a survey of the first 15 years of this literature.

This prediction persists in many extensions and modifications of the model. The positive correlation of tax rates and the population is one of the clearest predictions of models of tax competition, perhaps more general than any prediction on the shape of jurisdictions' fiscal reaction functions.

The basic tax competition model also provides some predictions about tax rates across cities, when the population distribution within these cities differs. Two simple comparative-static exercises can be considered. First, consider a city with n identical municipalities. In a symmetric equilibrium, all these municipalities will levy the same tax rate. The basic tax competition model predicts that this tax rate should decrease with the number n of municipalities.¹⁶ Fiscal equalization schemes redistributing tax revenue as a function of municipalities' *per capita* tax base will attenuate the rate-lowering effect of horizontal tax competition (see Köthenbürger, 2002; Bucovetsky and Smart, 2006). This attenuating effect of fiscal equalization has been confirmed in empirical research, for example, by Buettner (2006) and Egger et al. (2010). Second, consider the extent of asymmetry within a city. If there are only two municipalities within the city, the basic tax competition model predicts that the average tax rate within the city should increase with the degree of asymmetry—the population share of the larger municipality. This prediction again contrasts with the predictions from alternative models. New economic geography models suggest that tax differences are due to the ability of large jurisdictions to extract the rents from agglomeration economies. Such models imply that greater asymmetry will be associated with higher taxes in the larger municipality. Models with (some) population mobility lead to similar conclusions. In the Kanbur and Keen (1993) model of cross-border shopping, more asymmetry leads to lower tax rates in each municipality, and to a lower average tax rate for the whole city.¹⁷

The mechanism behind these conclusions from the basic tax competition model is quite straightforward. Suppose that a municipality's output is a quadratic function of its employment of mobile capital. If capital were perfectly mobile among municipalities, the quantity of capital attracted to a municipality would be proportional to the difference between the average unit tax rate on capital in the city, and the tax rate in the municipality. If municipalities differ only in size, then this relation between taxes and capital employment is exactly the same in each municipality, and can be written as

$$k_i = \bar{k} + \beta(\bar{t} - t_i), \quad (17.4)$$

where k_i is the capital employed per resident in municipality i , $\bar{k}$ is the capital available per person in the city, t_i is the tax rate in the municipality, $\bar{t}$ is the average tax rate in the city,

¹⁶ This literature assumes n to be exogenously given. For models of endogenous local jurisdiction formation, see, for example, Henderson (1985), Alesina et al. (2004), or Gordon and Knight (2009).

¹⁷ This model has been extended to analyze income tax competition when people are less than perfectly mobile by Gabszewicz et al. (2013).

and $1/\beta$ is the coefficient on the quadratic term in the production function. However, the average tax rate $\bar{t}$ in the city itself will depend on municipality i 's own tax rate. The larger the population share in the municipality, the stronger the influence of the municipality's own t_i on $\bar{t}$. So, tax reductions have a smaller impact on the capital employment per person in larger municipalities, since they have a larger absolute effect on the city's average tax rate. This differential impact underlies the positive correlation within a city between the municipal population and the equilibrium municipal tax rate.

The implications of the basic tax competition model for tax rate differences across cities are derived in [Bucovetsky \(2009\)](#). In addition to assuming a quadratic production function, he assumes that the marginal rate of substitution between the tax-financed local consumption good and the numéraire is constant. So, residents of a municipality seek to maximize $x_i + (1 + \varepsilon)t_i k_i$ subject to Equation (17.4), where x_i is private consumption $f(k_i) - (r + t_i)k_i + r\bar{k}$ (with r the city's net return to capital and $f(\cdot)$ the quadratic production function), and $\varepsilon > 0$ measures the premium placed on public consumption. The Nash equilibrium tax rate in a municipality can be expressed as a function of the average tax rate in the city and the share of the population in the municipality. Not only is the equilibrium tax rate an increasing function of the municipality's population, it is also a convex function of the population.

Because of this convexity, the overall average level of municipal tax rates within a city depends on the concentration of the population among municipalities. A "concentration index" for the population, similar (but not identical) to measures of concentration used in industrial organization, determines the city's average tax rate. Any movement of the population from a smaller municipality to a larger municipality within the same city must raise the equilibrium tax rate in the city.

In this framework, the share of the population of the largest municipality in a city plays an important role. Conditional on the largest municipality's share of the city's population, the average tax rate in the city still depends on the distribution of the population among the remaining municipalities. This rate will be highest when the population of these other municipalities is most concentrated and lowest when the population is least concentrated.¹⁸ However, as [Figure 17.10](#) illustrates, for a given share of the population in the largest municipality, the average tax rate is not too sensitive to the population distribution in the remaining municipalities. For comparison, [Figure 17.10](#) also shows the negative relationship between the population share of the largest jurisdictions and the average tax rates predicted by the [Kanbur and Keen \(1993\)](#) model described above.

¹⁸ Here "most concentrated" means that there are k other municipalities each with the same share s_1 of the population as the largest municipality and one other municipality with a share $1 - ks_1$, where k is the largest integer less than or equal to $1/s_1$. "Least concentrated" means that there are n other municipalities, each with a share $\frac{1-s_1}{n}$ of the population, and $n \rightarrow \infty$.

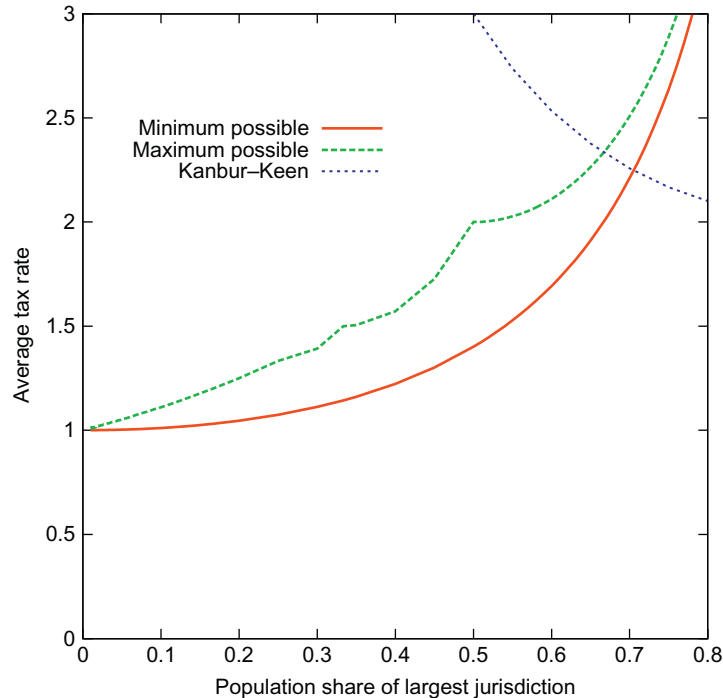


Figure 17.10 Average tax rate as a function of the largest jurisdiction's population share in the basic tax competition model. The average tax rate in a city is particularly sensitive to the population share of the largest municipality but is not too sensitive to the population distribution in the remaining municipalities.

In the basic tax competition model, tax revenue is used exclusively to finance a public consumption good. [Peralta and van Ypersele \(2005\)](#) assume instead that the revenue is used to make cash payments to residents.¹⁹ They also allow for municipalities to differ in two attributes: population, and capital endowment per person. They show that a ranking of municipalities by population similar to that in the earlier literature can be made.²⁰ Given quadratic technology, and assuming that taxes are strategic complements, the absolute value of a municipality's capital tax rate increases with its population, with the capital endowment per person being held constant.²¹ The mechanism here is similar to that in the basic model: taxes are levied solely to influence the city's terms of trade, and larger municipalities have a greater influence on them. It also remains true that smaller municipalities do better than large municipalities. The additional dimension of heterogeneity

¹⁹ Alternatively, the public consumption good is assumed to be a perfect substitute for the private good.

²⁰ [Peralta and van Ypersele \(2005\)](#), Proposition 4, p. 268.

²¹ Recent empirical research suggests that strategic complementarity of local tax rates may not be an innocuous assumption ([Parchet, 2014](#)).

also influences tax rates. If the population is held constant, municipalities with higher *per capita* endowments have lower tax rates. Here “lower” does not mean “lower in absolute value.” Municipalities with the highest endowment *per capita* will subsidize capital in equilibrium.

In [Pieretti and Zanaj \(2011\)](#), there is also no public consumption good. But in their model, municipal governments are leviathans, seeking to maximize net tax collections. There are several other features of the Pieretti–Zanaj model that differentiate it from the basic model. One attractive feature is that municipal governments provide a public intermediate good, which benefits investors; this feature makes the model more relevant to urban location models, in which residents’ location decisions may be influenced by the local public sector. So, municipal governments here seek to maximize revenue collected from source-based taxes on investment, net of the cost of public intermediate investments which attract that investment.

There are two other novel features. Capital owners within each municipality prefer to invest at home, rather than in other municipalities. They are also heterogeneous in terms of the strength of this preference. A type x investor incurs a cost of αx from investment outside his/her home municipality, the parameter α measuring the strength of the home-biased preference. The idiosyncratic attachment-to-home parameter x is assumed to be uniformly distributed over $[0,1]$ in each municipality. Moreover, the assumed production technology is different. The return on investment in any municipality is constant, rather than a decreasing function of the level of investment in that municipality.²² This constancy of the return on investment eliminates the channel through which population influences tax rates in the basic model. In the basic model, municipal tax rates influence the gross return on investment in the city, and the greater influence of larger municipalities’ taxes on that return explains why they set higher tax rates in equilibrium. In [Pieretti and Zanaj \(2011\)](#), this citywide gross return is fixed.

These modifications to the basic model weaken the positive relationship between population and tax rates. Whether the smaller municipality levies the lower tax rate depends on the degree of capital mobility. When capital is less mobile (α is high), the smaller municipality will levy the lower tax rate. But if capital is sufficiently mobile, the smaller municipality will levy a higher tax rate, and will still be a capital importer because of the higher quality of its productive infrastructure.

The main theoretical findings on horizontal tax competition are summarized in the following results:

Result 17.3.1. *In the basic model of horizontal tax competition, if a city contains n identical municipalities, then the municipalities’ equilibrium tax rates are a decreasing function of n .*

²² This is as in [Marceau et al. \(2010\)](#), or [Köthenbürger and Lockwood \(2010\)](#), for example.

Result 17.3.2. *In the basic model of horizontal tax competition, tax rates within a city will be positively correlated with municipal populations.*

Result 17.3.3. *In the basic model of horizontal tax competition, increased concentration of the population among the municipalities of a city increases the average tax rate in the city.*

Result 17.3.4. *If local public expenditure increases the return on investment, tax rates within a city may be negatively correlated with municipal populations, provided that the return on investment is not sensitive to the level of investment in the municipality.*

17.3.1.2 Empirical studies

Although many empirical studies of tax competition have provided evidence of the relationship between a municipality's population and its tax rate, in many instances, this relationship was not the focus of the study.²³ Estimating the slope of municipalities' reaction functions (one municipality's tax rate as a function of another's) is an active research area. The sign of these slopes is not directly relevant for the theoretical results emphasized here.²⁴ But a municipality's population is often used as a regressor in this empirical work, providing some evidence for how tax rates vary with the population within a given city.

Hauptmeier et al. (2012) offer fairly strong empirical support for the more conventional tax competition outcome, although in a framework that is closer to that of Pieretti and Zanaj (2011). They estimate the determinants of local tax rates and of local public input provision (roads) for a sample of municipalities in Baden-Württemberg in four different time periods. The coefficient on the population in the tax regression is positive and significant in nearly all the estimated equations. It is the effect of the population on reaction functions that is being estimated here, while the theoretical result of Pieretti and Zanaj (2011) pertains to the effect of the population on equilibrium tax rates. Nonetheless, under relatively weak conditions, a jurisdiction which wants to set a higher tax rate than another, when all jurisdictions levy the same tax rates, will be the one choosing the higher tax rate in equilibrium.

²³ For example, population is one of the regressors in the regressions run in Buettner (2006) in his study of the effects of fiscal equalization programs on tax setting, but the coefficient on the population is not reported.

²⁴ The sign of these slopes is indeterminate in the basic tax competition model. The results in Wilson (1991) for two-municipality cities hold regardless of the sign of these slopes. The stronger assumptions in Bucovetsky's (2009) extension to more than two municipalities imply that reaction curves must slope up, but there is no suggestion that this positive slope is necessary for the results. For an application to within-city tax reaction functions, see Brueckner and Saavedra (2001). For a promising new way to identify tax competition, based not on reaction functions but on estimable differences between desired and equilibrium tax rates, see Eugster and Parchet (2014).

The studies summarized in [Table 17.5](#) all provide estimates of the effect of a jurisdiction's population on the level of some fiscal variable in that jurisdiction. With one exception, the fiscal variable is a tax rate: business property tax rates, income tax rates, or excise tax rates on tobacco or gasoline. The one exception in the table is the article by [Solé-Ollé \(2006\)](#), in which a jurisdiction's total public expenditure is the dependent variable. The theory would predict a positive coefficient here (only) if a jurisdiction's tax base were an inelastic function of the tax rate.

With one exception, the articles cited in the table estimate reaction functions. Typically the lagged value of some weighted average of neighboring jurisdictions' tax rates is the independent variable of interest. The exception is the article by [Egger et al. \(2010\)](#), where the main focus is on the effect of equalization grant rules on a jurisdiction's tax rate. This is the one study that estimates the reduced form presented in the previous section: the equilibrium tax rates in municipalities as functions of exogenous variables. Because of fixed effects, the dependent variable is the change in a jurisdiction's tax rate, and the change in that jurisdiction's population is one of the explanatory variables. The coefficient on this variable is close to zero, varies in sign across regressions, and is not significantly different from zero in any of the reported results.

There is less empirical work so far on the relation between population concentration within a city and the average tax rate in a city. One article that does examine this relation deals as well with vertical tax competition and so is discussed in [Section 17.3.2](#).

The main empirical findings on horizontal tax competition among asymmetric jurisdictions in [Table 17.5](#) do not mirror the prediction of most theoretical models:

Result 17.3.5. *Empirical estimates provide conflicting evidence for the effect of population size on jurisdictions' tax rates.*

Table 17.5 Empirical evidence for the effect of population size on local tax rates

Article	Dependent variable(s)	Country/state	Year(s)	No. of jurisdictions	Sign of population
Allers and Elhorst (2005)	Property tax	Netherlands	2002	496	+
Bordignon et al. (2003)	Business property tax	Milan Province	1996	143	—
Devereux et al. (2007a)	Tobacco, gasoline	USA	1977–1997	48	Mixed
Edmark and Ågren (2008)	Income tax	Sweden	1993–2006	283	—
Egger et al. (2010)	Change in business tax	Lower Saxony	1998–2004	440	0
Hauptmeier et al. (2012)	Business tax	Baden–Württemberg	1998–2004	1100	+
Solé-Ollé (2006)	Expenditure	Spain	1999	2610	Convex

17.3.2 Vertical tax competition

While the basic model of tax competition among same-level governments implies that tax rates are set inefficiently low, that conclusion may be reversed when different levels of government share a common tax base and set their tax rates noncooperatively. Suppose that there is a higher-level city government that can levy its own taxes, on top of those set by lower-level municipal governments. This vertical interaction yields an additional externality. If the city contains N identical municipalities, the extent of this vertical tax externality, like the horizontal externality, increases with N . A single municipality, in choosing to raise its tax rate, lowers the tax base of the higher-level city government. Residents of the municipality bear a share $1/N$ of the costs of that tax base reduction, and so are less inclined to internalize the costs of this effect, the larger is N .²⁵ Because of these offsetting effects, it is not immediately clear whether increased decentralization leads to higher or lower taxes. Keen and Kotsogiannis (2004) analyze this sort of model. While the effects of increased decentralization on tax rates cannot be signed, the effect on overall welfare can. Since the vertical and horizontal externalities both increase with the number of municipal governments, equilibrium welfare of residents must fall. The Keen–Kotsogiannis model also extends the basic tax competition model by allowing for a variable total supply of capital (to the city), as an increasing function of the net return to capital. Without this extension, there would be no vertical externality, since the city's tax base would not be affected by municipal tax rates.

A natural extension of the Keen–Kotsogiannis model is to model explicitly competition among cities. The article by Wrede (1997) was one of the first works to attempt this extension, although vertical tax competition was assumed away in this article. More recent work includes the work of Hoyt (2001) and Breuillé and Zanaj (2013). Breuillé and Zanaj maintain the assumption of fixed aggregate capital supplies prevalent in the horizontal tax competition literature. But with several cities, vertical externalities will occur even with this fixity: a tax increase in one of the municipalities reduces capital supply to the whole city. With quadratic production functions, and preferences separable in the public consumption goods provided by each level of government, Breuillé and Zanaj show that a merger of any municipalities must (1) raise tax rates set by each city (including those not party to the merger), (2) lower all municipal tax rates, and (3) increase all combined (city plus municipal) tax rates.

Result 17.3.6. *With vertical externalities, tax rates may increase with the number of municipalities in a city if the city's overall capital supply function is upward sloping.*

Brühlhart and Jametti (2006) modify the Keen–Kotsogiannis model to a form in which the relative importance of vertical and horizontal effects can be tested empirically. In their

²⁵ Keen and Kotsogiannis (2002), p. 366.

theoretical model, it is assumed that each city contains N identical municipalities. The marginal payoff from a tax increase in municipality i can be decomposed between two terms, one due to horizontal externalities and the other to vertical externalities. In that model, a positive correlation between the number of municipalities and the tax rates in the municipalities indicates that vertical externalities are more important.

Brülhart and Jametti (2006) estimate the relation between a municipality's population share and its tax rates using a panel of Swiss municipalities. The sample is divided between the set of municipalities in which decisions must be approved directly at a public meeting open to all citizens, and those in which, instead of open meetings, local government decisions must be approved by a referendum. The first sample corresponds to municipalities in which decision making is most likely made by some representative citizen; referenda give considerable power to local government officials through their control of the agenda. Brülhart and Jametti find a significant negative relationship between a municipality's population share and its tax rates for the first (public meeting) subsample, indicating that vertical externalities dominate horizontal ones. The relationship is also negative for the second (referendum) subsample, although the coefficient here tends not to be significant.

Result 17.3.7. *Evidence from Swiss municipalities indicates that at the local level vertical tax externalities may be as relevant as horizontal tax externalities.*

17.3.3 Voting

One of the major weaknesses of the basic model in its applicability to urban taxation is the assumption that people are immobile. We now turn to literature that models the relation between the municipal population and tax rates when people are mobile, differ in income, and vote on tax policy. In this subsection, we focus on the interaction of the voting equilibrium within municipalities and the sorting equilibrium across municipalities. See also the chapter by Ross and Yinger (1999) on the early contributions. Section 17.4 focuses in more detail on population sorting across municipalities. Within a municipality, the shape of the distribution of income across the population will determine the progressivity of the tax system. A widely used, tractable model of voting on income tax schedules is that of Romer (1975), Roberts (1977), and Meltzer and Richard (1981), in which self-interested voters choose a flat income tax, the proceeds of which are distributed equally (thus implying a progressive effect of the system), and in which the efficiency loss due to this distortionary tax is a quadratic function of the tax rate. In such a model, the Condorcet winner among tax rates chosen by majority rule is proportional to the ratio of the median income to the mean income.

These models can be used to explain the variation of tax rates across municipalities, if municipalities differ in population, and in the distribution of income over that population.

Suppose, for example, that the distribution of income in some municipality was a truncated Pareto distribution over some interval (L, H) . A property of that distribution is that the ratio of median income to mean income in the municipality is a decreasing function of the ratio of the lowest to the highest income L/H , independent of the “scale” parameter L . This means that if a municipality contained only people whose incomes fell in some segment (L, H) of the overall metropolitan income distribution, then the median–mean ratio would have to be very close to 1, if the segment were small enough.²⁶ In other words, the actual size of a municipality may affect the tax rate chosen: small municipalities will not vote to levy high income taxes.

Suppose that people vote on tax schedules taking the population composition as given.²⁷ That is, either voters are myopic, and ignore the effect of their own choices on mobility, or voting takes place after people have made their location choices. When will a sorting equilibrium arise in which each municipality contains a disjoint segment of the income distribution? If municipalities are numbered in increasing order of their income, so that municipality j contains a slice (L_j, H_j) of the income distribution, with $H_j = L_{j+1}$, then a necessary condition for the population allocation to be an equilibrium is that the person with income $H_j = L_{j+1}$ is indifferent between municipalities j and $j + 1$. If people care only about their net income, and if municipalities are allowed to charge a flat admission fee p_j to each resident, the payoff from choosing to reside in municipality j for a resident with (exogenous) income γ is $\gamma(1 - t_j) + t_j \bar{\gamma}_j - p_j$ if the income tax rate in the municipality is t_j and if income tax proceeds are distributed equally to all residents. If the equilibrium is stratified, and if γ is the highest income level in municipality j , then

$$\gamma(1 - t_j) + t_j \bar{\gamma}_j - p_j - \gamma(1 - t_{j+1}) - t_{j+1} \bar{\gamma}_{j+1} + p_{j+1} = 0. \quad (17.5)$$

In order for a stratified equilibrium to exist, not only must Equation (17.5) hold as an equality for people with income $\gamma = H_j = L_{j+1}$, but the left side of the equation must be decreasing in γ near $\gamma = H_j = L_{j+1}$: people with income below H_j must prefer strictly to live in municipality j , and people with income higher than L_{j+1} must prefer strictly municipality $j + 1$.

Therefore, a necessary condition for the existence of a sorting equilibrium in this sort of model is that the tax rate be lower the higher is the income in the municipality.

Underlying Equation (17.5) is the assumption that all entrants to a municipality pay the same entry fee p_j regardless of income. But the result can be generalized: if entry to municipalities is rationed by differences in (unit) housing prices, this necessary condition still holds, provided that the income elasticity of the demand for housing is less than 1.

²⁶ And if L were bounded away from 0.

²⁷ This is, for example, assumed in Calabrese et al. (2006) and Epple et al. (2012) as discussed in Section 17.4.1.1.

Hansen and Kessler (2001) reconcile these two necessary conditions for the existence of sorting equilibrium with voting on tax schedules that (1) the tax rate in a municipality depends only on the shape of the income distribution within the municipality, and (2) the income tax rate chosen in each municipality must be lower the higher is the segment of the income distribution of people choosing to live in the municipality.

The characterization above shows that for a sorting equilibrium to exist, the ratio of median income to mean income needs to increase as we move to higher segments of the income distribution. That could not happen, for example, if the overall income distribution for the city were uniform. But if the distribution were a Pareto distribution, the discussion above shows that this ratio will increase if (and only if) the ratio of the lowest to highest income increases as we move to higher segments of the income distribution. Now suppose that the upper bound for income for the whole city were some finite Y . Then if one municipality were sufficiently small, the ratio of the lowest to highest income L/H would have to be close to 1 if the municipality contained a segment of the income distribution. So if the city contained only two municipalities, one much larger than the other, and if the income distribution for the whole city were a truncated Pareto distribution over some interval $[y_0, Y]$, then the tax rate would have to be lower in the smaller municipality in any sorting equilibrium.

Hansen and Kessler (2001) generalize this result.²⁸ The restriction to the Pareto distribution is not necessary. As long as the overall income distribution has finite support, the ratio of the median income to the mean income in any one municipality must approach 1 if the municipality contains a small enough segment of the income distribution. They can therefore show that if a city is divided into two municipalities, then a sorting equilibrium will exist if population asymmetries between municipalities are large enough. And, in this case of large disparities, the stratification must involve the smaller municipality having the lower taxes, and the higher segment of the income distribution.

Now, stratification is not the only possible equilibrium. A completely symmetric distribution, in which equal shares of all income groups move into each municipality, will always be an equilibrium. Hansen and Kessler (2001, p. 1109) show that there are cases in which no sorting equilibrium exists. Hence, asymmetries in municipal size may lead to symmetries in taxes: only if there are large enough size differences can a heterogeneous equilibrium exist.²⁹

The main finding on voting is summarized in the following result:

²⁸ They need to assume that the mean income for the whole city exceeds the median income, and that the income distribution for the whole city is unimodal. This latter requirement ensures that the mean income exceeds the median income in the richest municipality, so that all municipalities have positive tax rates in equilibrium.

²⁹ For this model, Equation (17.5) shows that if there are any differences at all in tax rates across municipalities, then there must be complete stratification.

Result 17.3.8. *Voting within municipalities implies that municipal income should be negatively correlated with the municipal population if municipalities are stratified by income.*

17.3.4 Central municipalities and suburbs

The distinction between the center and the suburbs seems important, and has not been considered much in the theoretical literature on tax competition. There are a few recent articles that have emphasized this distinction. The models in these articles are quite specific, and the conclusions appear quite sensitive to the modeling assumptions. These articles certainly represent an important step in the right direction. But further work seems needed in order to establish plausible, tractable theories of the effect of urban structure on municipal tax policy.

One such model is that of [Janeba and Osterloh \(2013\)](#). They show how tax competition among cities may affect the central municipality of each city more than the suburbs. In their model, as the total number of cities becomes large, the tax rates set by each municipality within each city shrink. But the suburban tax rates approach a positive asymptote, whereas central tax rates approach zero as the number of cities grows.

There are no differences among cities in this model, but there are differences within each city. Each city contains a single central municipality with a share $1 - s$ of the total city population, and m suburbs, each with a share s/m of the population.

The other differences in the model between central and suburban municipalities help drive the results here. Capital is assumed to be mobile among cities and within cities, but it is assumed to be more mobile within cities. Decision making is sequential. Central governments set their tax rates (simultaneously) first. Next, owners of capital decide how to allocate their capital among different cities. Then suburban municipalities choose their tax rates, at which time the quantity of capital within the city has already been fixed. Finally, capital owners decide on the allocation of capital within cities.

These differences in timing are crucial. The authors use evidence from a survey of local decision makers as the basis for their assumption.³⁰ The survey asked mayors which jurisdictions they perceived as providing the most competition (with their own jurisdiction) for business. The responses show that mayors of less populous jurisdictions tended to regard other jurisdictions within the same state as their strongest competitors, while mayors of more populous jurisdictions perceived additional competition as arising from jurisdictions outside the state or country.

Labor is supplied inelastically, and people are immobile. Each municipality finances a local public consumption good, with a head tax and a source-based capital tax. The availability of the head tax means that public good supply is efficient. Capital taxes are used to

³⁰ One thousand one hundred eight mayors in Baden-Württemberg were contacted, of whom 714 responded.

influence the net return to capital: central governments seek to influence the “national” net rate of return, and suburban governments see to influence the return within the city.

Capital taxes are positive in equilibrium, despite the symmetry within and across cities. This incentive to use positive capital taxes to lower the return to capital comes from an assumption of asymmetries *within* municipalities. While the pattern of capital and labor ownership is identical across cities (and across municipalities within a city), it differs within each municipality. As seems realistic, the distribution of the ratio between capital endowment and labor endowment is assumed to be asymmetric, with the median less than the mean. Therefore, the median voter gains from a fall in the net return to capital, even though there is no net inflow or outflow of capital to the “world” as a whole.

Assuming that output per person is a quadratic function of the capital–labor ratio enables the study authors to derive closed-form expressions for the central and suburban capital tax rates in a symmetric equilibrium. These tax rates are both positive. They depend on the number n of cities, the number m of suburbs within each city, and the fraction s of the population in the suburbs of each city. As the number n of cities increases, capital tax rates everywhere decrease, but they decrease more rapidly in the central municipalities.³¹ Since capital taxation is not necessary for the public sector here, owing to the availability of the nondistortionary tax, the conventional result in single-tiered tax competition is that tax rates should approach zero as the number of identical municipalities becomes large. Here that result continues to hold for centers, but it will not hold for suburbs. Suburban tax policy is directed at affecting the return to capital within the city, *after* the allocation to capital across cities has been determined. As $n \rightarrow \infty$, the suburban tax rate approaches the positive asymptote, which is proportional to $\frac{\hat{e}}{1 + \hat{e}m - s^2}$, where $\hat{e}$ is the (common) median value of the ratio of labor endowment to capital endowment, relative to the mean, a tax rate that decreases with the number m of suburbs in the city, and with the share $1 - s$ of the population in the center. Because the central voters make their decisions before suburban voters, (only) the central voters worry about the effect of their tax increases on the overall supply of capital to the city. Even if the geography were symmetric within each city—one center and a single suburb with the same population in this model—the tax rate would be higher in the suburb if there is more than one city.

Kächelein (2014) develops a model similar to that of Janeba and Osterloh (2013), in that there are asymmetries within cities, and complete symmetry among different cities.³² However, fiscal policy takes place in a single stage, and there is no distinction between capital movement within and across cities. In Kächelein’s model, identifying different

³¹ This is Proposition 2 of Janeba and Osterloh (2013).

³² Braid (1996, 2005) has also developed a model of tax competition with commuting: residential location is fixed, but people may commute within a city, the supply of capital to the city is perfectly elastic, and some of the land in a municipality may be owned by nonresidents. He assumes that all municipalities in a city are identical, and so does not address the asymmetries which are the focus of this subsection. Nonetheless, these articles represent an important step in addressing tax competition which is truly urban in nature.

municipalities as “central” and “suburban” is less obvious. In this model, workers may commute within a city. In equilibrium, workers commute from the larger municipality to the smaller municipality, so the larger municipality is best identified as an aggregate of suburbs, rather than as the center.

In the base case, municipalities have a single tax instrument, a source-based tax on capital employed within the municipality. The revenue from this tax is used to finance a publicly provided consumption good.³³ It is assumed that the number of cities is large, so the world return to capital is unaffected by changes in any single municipality’s tax rate. Thus, absent the commuting, each municipality would levy the same tax rate.

However, here there are only two municipalities within each metropolitan area, and workers can commute only within the city. Therefore, each municipality’s capital tax rate will affect the wage rate in that municipality, and in the other municipality in the city. Municipalities are also assumed to differ in their population—but not in their endowment of a third, immobile, factor, land. In the absence of any tax differences, some residents of the more populous municipality would choose to work in the other municipality. That means that the smaller municipality can export some of the burden of its source-based capital tax onto workers who commute from the larger municipality. In the article, it is shown that this tax-exporting effect yields somewhat similar implications for the implications of population asymmetries as those in the basic tax competition model. In particular (Proposition 2), residents of the smaller municipality will be better off in equilibrium, and the smaller municipality will levy lower tax rates (Proposition 1). In an extension of the model, Kächelein shows that the availability of a source-based wage tax does not change the basic results of the model: municipalities still use capital taxation, and larger municipalities tend to rely more on wage taxation and less on capital taxation. This latter result is a prediction about the effect of the relative population on the tax mix, which is not present in most other models of taxation and asymmetry.

In Gagné et al. (2013), production is not restricted to the central business district (CBD). However, the geographic center of the city—which need not coincide exactly with the central municipality—is assumed to be more productive. Workers’ productivity is fixed, and is at least as high in the city’s CBD as it is anywhere else in the city.³⁴ The city is (exogenously) divided into $m + 1$ jurisdictions, a central municipality, and m identical suburbs. The city is “one-dimensional” in that it consists of m rays through the CBD. People living along any of the rays at a distance less than b from the CBD are residents of the central municipality. Those living further than b from the CBD reside in a suburb: each ray contains a different suburb. (Both m and b are taken as exogenous.)

From the point of view of efficiency, there are two offsetting costs determining the pattern of location and employment. As mentioned, workers are at least as productive in

³³ That is an imperfect substitute for the private good.

³⁴ Production is assumed not to use any land, so all production in each location takes place at a single point.

the CBD, at the geographic center of the city. But commuting to work is costly. Gagné et al. (2013) also assume that the costs of the local public sector depend on population size. In particular, the *per capita* cost of each municipality's public sector is a U-shaped function of the population served. These population effects on the cost of the local public sector mean that the location of municipal borders matters, separately from the location of municipalities' employment. Since productivity throughout the city does not vary—except in the CBD—commuting costs within a suburb are minimized by locating employment at the midpoint of the municipality's employment region. This region may differ from the political boundaries, as some suburban residents may choose to work in the center.

Therefore, each suburb contains an employment point at a distance $(y + B)/2$ from the CBD, where B is the radius of the city and y is the location of the central workers who reside at the greatest distance from their workplace.

From a social planner's viewpoint, residential location can be decoupled from employment location. The cost function for the public sector implies that a given population should be divided evenly among all the occupied municipalities.³⁵ If there were no cost advantage to central employment, minimization of commuting costs would require $y = B/3$. Equating the population of all the municipalities requires that the city's radius b be a fraction $1/(m + 1)$ of the distance to the edge of the city. Thus, from the planner's perspective, $y \geq b$ if and only if $m \geq 2$. With three or more suburbs, it is optimal for some workers to commute from the suburbs to the center. Having a positive productivity advantage in the CBD strengthens this effect.

Under decentralization, municipal governments choose fiscal policies to maximize residents' incomes. As in much of the literature, Gagné et al. (2013) simplify the choice of the maximand by having residents commit to location choices before the local public sector is determined, anticipating the equilibrium choices that will be made. Further, they assume that the land rents in each municipality are divided equally among all residents of the municipality. The size of each local public sector is assumed to be fixed. The fiscal choice made in each municipality is how to divide the cost of the local public sector. Taxes may be levied on those who live in the municipality, and on firms located there. Firms operate under constant returns to scale; free entry and the fact that labor is the only input to production imply that the incidence of the tax on firms falls entirely on people who work in the municipality.³⁶ When $b \neq y$, a municipality's workers are not the same as its residents.

³⁵ Since the total cost of the public sector of a municipality of population P is assumed to be $F + \alpha P^2$, these costs decrease with the number $m + 1$ of municipalities if and only if $m + 1 \leq \sqrt{\frac{\alpha}{F}}L$, where L is the total population of the city.

³⁶ Recall that production does not use land in this model.

Any difference in the tax rate levied on firms in the center and on those in the suburbs will be distortionary. In the absence of tax differences, workers (and firms) locate so as to maximize net output in the city, minus commuting costs. A higher tax on central firms than on suburban firms must imply an inefficient pattern of production, with not enough production in the center.

Gagné et al. (2013) show that a municipality will levy a positive tax rate on firms if there are people from elsewhere who choose to work in the municipality. It can export some of the tax burden.

But they also derive two more specific and surprising results. The only possible commuting is from the suburbs to the center. Depending on commuting costs β , the size B of the city, the radius b of the center, and the cost advantage $E \geq 0$ of the CBD, three cases arise: all production takes place in the center if $b \geq 5B/3 - 2E/\beta$, no commuting takes place in equilibrium ($y = b$) if $5B/3 - 2E/\beta > b \geq B/3 + 2E/3\beta$, and otherwise there is some commuting from the suburbs to the center. The center will therefore choose a positive tax on firms, except in the second case ($b = y < B$), in which case it levies a tax of zero.

The second specific result is that the suburbs choose not to tax firms, nor to subsidize them, in equilibrium. The tax rate T levied by any suburb and the central tax rate T_0 determine the boundary y between the employment zone of that suburb and that of the center. The star-shaped nature of the city ensures that there is no interaction here between different suburbs. Suburban governments choose their tax rate T so as to maximize total output produced by residents of the suburb, minus commuting costs of residents, minus taxes paid to the central government by suburban residents who commute to the CBD. A tax rate of zero turns out to maximize this net output.

These two specific results imply that there will be too little commuting to the center. The model yields a closed-form solution for the tax rate on firms in the center:

$$T_0 = \frac{E}{2} + \frac{\beta(B - 3b)}{4}. \quad (17.6)$$

Equation (17.6) implies a relationship between the population of the center and its tax rate. Since b is the radius of the center, and B is the radius of the city, Equation (17.6) implies that the source-based tax rate T_0 in the center will decline with the center's share of the city's population. Other things being equal, an increase in the center's population means that there are fewer suburban commuters to whom to shift the tax burden, so the center's optimal T_0 declines. Since suburbs set a source-based tax rate T of zero, regardless of their size, the model of Gagné et al. implies that more concentration of the population in the central jurisdiction leads to *lower* source-based tax rates in the city.

The maximum commuting distance y to the CBD is determined by the equality, for the marginal worker, of net-of-tax earnings in the CBD, and in the suburban employment center, located halfway between the employment boundary y and the outer boundary of the city. Therefore, y must satisfy

$$\beta\gamma + T_0 = \beta\left(\frac{B+\gamma}{2} - \gamma\right) + E + T. \quad (17.7)$$

Given Equation (17.6) and this definition of γ , the fact that $T = 0$ in equilibrium implies that

$$\gamma = \frac{1}{6}\left(B + 3b + 2\frac{E}{\beta}\right). \quad (17.8)$$

When there are no productivity differences among locations ($E = 0$), Equation (17.8) implies that $\gamma = \frac{B}{6} + \frac{b}{2}$, which must be less than the efficient radius of employment $\gamma^* = B/3$ mentioned above. The center's shifting of the tax burden onto commuters leads to too little CBD employment. This result continues to hold when the productivity advantage of the CBD is strictly positive.

The results in [Janeba and Osterloh \(2013\)](#) imply the following:

Result 17.3.9. *If capital is mobile among cities and central tax rates are important in firms' location choices among cities, tax rates will be lower in the center.*

The results in [Gagné et al. \(2013\)](#) imply the following:

Result 17.3.10. *In a single-city model in which residents can commute, tax rates on capital will be higher in the center if the direction of commuting goes from the suburbs to center. These tax rate differences imply that too few of a city's workers will work in the center.*

17.4. TAXATION AND URBAN POPULATION SORTING

17.4.1 Tax-induced urban population sorting: theory

17.4.1.1 Aspatial general equilibrium models

Starting with [Ellickson \(1971\)](#) and [Westhoff \(1977\)](#), there is a long tradition of formally modeling fiscal decentralization *within* cities populated by heterogeneous agents in the spirit of [Tiebout \(1956\)](#). This literature on multijurisdiction models has almost entirely focused on local *property* taxation and has been comprehensively surveyed in earlier volumes of the handbook ([Ross and Yinger, 1999](#); [Epple and Nechyba, 2004](#)). We therefore limit ourselves to recalling the basic setup of these models and the associated main results in this section.

In all of these models, households that differ in income choose among a fixed number of local jurisdictions (municipalities). The residents of the local jurisdictions vote on the provision of a local public good that is financed by local property taxes (see also [Section 17.3.3](#)). Note that the local “public good” in these models is strictly speaking a “publicly provided private good” as it is both excludable and rivalrous in consumption. Local public budgets are balanced and local housing prices adjust to the local demand for housing. The models are aspatial in the sense that distances

between and within local jurisdictions are irrelevant. The literature focuses on asymmetric equilibria with different levels of property tax rates, public good provision, and housing prices across municipalities. Sorting depends on the nature of the public goods and housing. In multijurisdiction models with property taxes, households base their location decision on the after-tax price of housing. The property tax rate itself is therefore not a separate location characteristic. Each household faces in equilibrium a trade-off between public good provision and after-tax housing prices. For models with linear property taxes, sorting depends on the nature of the public goods and housing:

Result 17.4.1. *If the income elasticity of housing demand equals 1 and public goods are easily substituted by private goods (e.g., a pure monetary transfer), then rich households sort into municipalities with low public good provision and low after-tax housing prices. If public goods cannot be easily substituted by private goods, then rich households prefer municipalities with high public good provision and high after-tax housing prices.*

Note that there is no theoretical prediction about the relation between municipal income levels and the property tax rate itself.

The basis for calibrating and estimating multijurisdiction models is the version in [Epple and Platt \(1998\)](#), where households are heterogeneous in both income and tastes, leading to realistic incomplete sorting by incomes. In the original Epple–Platt model, rich households sort into municipalities with low public good provision and low after-tax housing prices. This contradicts the pattern typically observed in US cities. Empirical applications of the model (e.g., [Epple and Sieg, 1999](#)) therefore use a version of the model of [Epple and Platt \(1998\)](#) in which public goods are not easily substituted by private goods and rich households sort into municipalities with high public good provision and high after-tax housing prices. More recently, the basic models have been extended in several dimensions. In [Calabrese et al. \(2007\)](#), the local population can also vote for zoning in the form of minimum housing size requirements. Zoning aggravates income sorting and increases aggregate welfare, but with significant welfare losses for the poorer households relative to the equilibrium without zoning. [Epple and Ferreyra \(2008\)](#) use the model with peer effects to study the effect of school finance equalization and show that the generalized model is able to correctly predict the observed effects of the school finance reform. [Epple et al. \(2012\)](#) incorporate an overlapping generations model where older households without children are less interested in school quality than younger households with children. The resulting equilibrium sorting by both income and age can reduce inequality in educational outcomes compared with models with income sorting only.

[Calabrese et al. \(2012\)](#) study the welfare effects in a calibrated version of the Epple–Platt model with local property taxation. They find both a *per capita* welfare loss and a welfare loss by most of the population in the decentralized equilibrium with population

sorting compared with a centralized equilibrium with no sorting. The welfare loss is small (less than 0.1% of mean income) and is mainly (more than 99.5% of the welfare loss) shouldered by landowners through falling rents. This welfare loss is remarkable as the decentralized equilibrium with high public good provision in small elite jurisdictions is a Tiebout-type equilibrium. They attribute the inefficiency of the decentralized property-tax equilibrium to “a jurisdictional choice externality, where relatively poorer households free ride on richer households in suburbs by buying small houses to avoid taxes” (Calabrese et al., 2012, p. 1082). The efficient equilibrium features decentralized local head taxes and produces substantial welfare gains compared with both the decentralized and the centralized property tax equilibrium. Interestingly, the head-tax equilibrium produces larger differences in public good provision and more sorting than the decentralized property tax equilibrium, realizing the efficiencies typically associated with Tiebout-type equilibria.

Result 17.4.2. *If public goods cannot be easily substituted by private goods and when households differ in both incomes and tastes, rich households and households with a strong taste for public goods sort into municipalities with high public good provision and high after-tax housing prices. Such equilibria with decentralized property taxation are not more efficient than equilibria with uniform public good provision, centralized property taxation, and no population sorting. Equilibria with decentralized head taxation are substantially more efficient than equilibria with property taxation but imply more pronounced sorting of the population.*

This welfare analysis, however, does not consider the informational advantages of decentralized decision making (see, e.g., Kessler, 2014), and it abstracts from equity concerns.

We showed in Section 17.2 that property taxation is the principal local government revenue source in the United States and in Commonwealth countries. In most other higher-income countries, income taxation is the more important local revenue source. Multijurisdiction models with local income taxation have been studied by Goodspeed (1989), Schmidheiny (2006a,b), and Hodler and Schmidheiny (2006). These models are identical to property tax models in all basic assumptions except for the tax base. Studying income taxes introduces a severe technical difficulty into the formal models: in income tax models, tax rates and housing prices are two separate dimensions of location characteristics. Analyzing the sorting of heterogeneous agents across municipalities that differ in more than two dimensions (tax rate, housing price, public goods) is challenging and only produces nonambiguous sorting under restrictive additional assumptions about household preferences (see Gravel and Oddou, 2014). Apart from the technical difficulties, the established results for income tax models are very similar to the results with property taxes. As in property tax models, the literature focuses on asymmetric equilibria with different levels of income tax rates, public good provision, and housing prices across municipalities. Sorting depends on the nature of the public goods and housing. In

multijurisdiction models with *income* taxes, each household in equilibrium faces three bilateral trade-offs between public good provision, housing prices, and income tax rates. This leads to a rich possible set of equilibrium configurations depending on the nature of the public goods and housing. [Goodspeed \(1989\)](#) establishes the following empirically relevant cases:

Result 17.4.3. *If the income elasticity of housing demand equals 1 and public goods are easily substituted by private goods (e.g., a pure monetary transfer), then rich households sort into municipalities with low income tax rates, high housing prices, and low public good provision. If public goods cannot easily be substituted by private goods, then rich households sort into municipalities with low income tax rates, high housing prices, and high public good provision.*

[Schmidheiny \(2006b\)](#) derives sufficient sorting conditions behind this result and extends it to incomplete sorting of households that differ in both incomes and preferences as in [Epple and Platt \(1998\)](#). [Schmidheiny \(2006a\)](#) introduces *progressive* income taxation as an alternative motive that predicts sorting of high-income households into low-tax municipalities.

Note that the housing market is essential in sustaining the empirically most relevant case in which high-income households locate in low-tax, high-public-good-provision municipalities. It is the high housing prices which prevent the low-income households from following the high-income households into the low-tax, high-public-good-provision locations.

Sorting of the population by income is a general phenomenon which is also observed in cities with uniform tax rates. The prime explanation is *social interactions* in various forms. See [Ioannides \(2013\)](#) for an extensive discussion of the theoretical and empirical literature. [Calabrese et al. \(2006\)](#) combine both social interactions and tax decentralization into a unified formal framework. The public good is modeled as expenditure on education and social interactions are modeled as peer group effects in education. This model leads to a rich set of possible equilibrium configurations. After calibrating the model (see also [Section 17.4.2.2](#)), they find that rich households sort into municipalities with high public good provision and high after-tax housing prices, as in the model without peer effects. However, different from calibrations of the basic model and in line with empirical observations, property taxes are *lower* in high-income municipalities than in low-income municipalities:

Result 17.4.4. *If there are peer effects in the production of educational quality, and the public provision of education cannot be easily substituted by private provision, rich households sort into municipalities with high housing prices, high public good provision, and low property tax rates.*

17.4.1.2 Asymmetry and centrality

There is no locational specificity in the models presented in the previous subsection: equilibria are unique only up to a permutation of the names of the municipalities.

The one feature in most of this work which distinguishes the central municipality from other (sub)urban municipalities is the size of the central municipality, which has a larger area or more housing units. So, if the central municipality has 40% of the housing stock, there will be one equilibrium in which the poorest 40% of the population live in the central municipality, but also another in which the poorest 60% live in the suburban municipalities.

Three articles by [de Bartolome and Ross \(2003, 2004, 2007\)](#) show that introducing locational heterogeneity into this sort of model can break the multiplicity of equilibria.³⁷ de Bartolome and Ross assume that workers commute to the CBD. The cost of commuting is assumed to be higher for high-income people than for low-income people. For tractability, demand for space is assumed not to vary with income. If the city comprises a single municipality, there would be a unique equilibrium location pattern in this model. People would sort by income, with the richest people living closest to the center of the city.

de Bartolome and Ross divide the area into two municipalities: a round central city with the CBD at its core and a ring-shaped suburban municipality that surrounds the central city. Each municipality provides a public output that is financed by a head tax. The income elasticity of demand for this public good is assumed to be positive. Hence, each jurisdiction provides the public output level preferred by the median-income resident of the jurisdiction. This heterogeneity of demand for the local public output generates multiple equilibria. There will still be an equilibrium in this model in which income throughout the metropolitan area *declines* monotonically with the distance from the center. There are two reasons why the rich will bid the most for land in the central municipality: they are willing to pay the most for accessibility to the CBD, and they prefer the public output provided in the CBD.

But there can be a “reverse equilibrium” as well. Suppose that the central city has a lower median income than the suburb. If that is the case, rich residents face a trade-off. The central city still offers better accessibility to the CBD, but it now offers a public output package which the rich like less than that provided by the higher-income suburb. If the income elasticity of public output demand is high relative to the income elasticity of commuting costs, then the second effect will outweigh the first. The theoretical argument above is presented in [de Bartolome and Ross \(2003\)](#). In subsequent work, de Bartolome and Ross constructed numerical examples which confirm that there may be equilibria in which the richest residents of a city locate in the suburban municipality in equilibrium. In their examples, a city consists of a circular central municipality surrounded by a single annular suburb. Each municipality chooses its public output level (financed by a head tax) by majority rule, and voters are myopic. In [de Bartolome and](#)

³⁷ [Epple et al. \(2010\)](#) also show that within-jurisdiction amenities such as the distance to the center can in principle be integrated into the multijurisdiction models presented in the previous subsection. However, they do not study the properties of the equilibrium.

Ross (2004) there are two income classes in the city, and in de Bartolome and Ross (2007) the distribution of (exogenous) income is continuous. In each model, there must exist an equilibrium in which the highest stratum of the income distribution chooses to locate in the central municipality. But in each model, for some parameter values there also exists a second, stable, equilibrium, in which the richest people choose to locate in the suburb in equilibrium, because of the higher level of public output chosen there. So, it remains true that, within a given jurisdiction, income declines with the distance from the CBD. In this second type of equilibrium, the very poorest people locate in the furthest-out portion of the central municipality, and the very richest locate just beyond them, in the part of the suburb closest to the CBD.

The second type of equilibrium appears somewhat inefficient, in that the people who are willing to pay the most for accessibility end up located fairly far from the center, because of the self-fulfilling belief that central cities are poorest. de Bartolome and Ross (2007) show that aggregate welfare, added up over all residents, is higher in the first type of equilibrium, in which the poor locate in the suburb. However, this first type of equilibrium does not necessarily Pareto dominate the second. de Bartolome and Ross compute an example in which the poorest people are actually better off in the second type of equilibrium, in which they live in the central city.

The findings of de Bartolome and Ross (2003, 2004, 2007) can be summarized as follows:

Result 17.4.5. *Assuming that high-income households have a higher relative willingness to pay for centrality than low-income households, high-income households sort into the more central parts of a city, which is formed of a single jurisdiction. This equilibrium sorting can be sustained when the city is partitioned into a large center municipality and many small suburban municipalities. There also exist “reverse” equilibria in which high-income households sort into the (sub)urban municipalities far from the center.*

17.4.2 Tax-induced urban population sorting: Empirical studies

The effect of local taxes on the location decisions of individuals and firms can in principle be studied by relating individual or aggregate location choices to the local tax burden.³⁸

This typically observed negative relationship, however, cannot be interpreted as a causal effect because of an intrinsic reverse causality problem. A large local tax base of high-income households can lead to high tax returns even when the tax rates are relatively low. Municipalities with a large tax base are therefore able to balance their budgets with

³⁸ Kirchgässner and Pommerehne (1996) and Feld and Kirchgässner (2001), for example, study Swiss municipal data and show that high-income households are systematically located in cantons and municipalities with low income tax rates. Liebig et al. (2007) show that Swiss municipalities with high tax rates have higher emigration than municipalities with low tax rates. They also show a positive correlation between tax rate changes and emigration.

lower tax rates. Individual location decisions therefore affect local taxes—at least in the long term—through the local budget constraint and the political process behind it. This produces reverse causality, which is hard to overcome in nonexperimental studies. This section outlines four different approaches to overcome the endogeneity problem.³⁹

17.4.2.1 Estimation of individual location choice

The first empirical approach directly targets the location choice of individual households in a multinomial response framework. [Friedman \(1981\)](#) uses a conditional logit model to study the location choice of 682 households among nine residential areas in the San Francisco area. [Nechyba and Strauss \(1998\)](#) apply the same model to study the choice of over 22,000 households among six school districts in the suburbs of Philadelphia. Both studies show that high public expenditures (such as per pupil school spending) and low per-unit housing prices attract residents. However, these studies cannot estimate the effect of the property tax rate *per se*, because it affects people's decisions only through the after-tax housing price.

This approach seemingly circumvents the endogeneity problem because, from the perspective of a single household, the community characteristics can be taken as given. However, the local tax rates are still a function of all individual location decisions. [Schmidheiny \(2006a\)](#) therefore focuses on the location choice of households which are moving in a given year. *Movers* are a relatively small share of the whole population, while the equilibrium tax rates are mainly driven by the large share of *stayers*. [Schmidheiny \(2006a\)](#) estimates the location choice of the universe of households that moved in 1997 from the central municipality to any other municipality in the Basel urban area in Switzerland. He starts from the following indirect utility function:

$$V_{ni} = \alpha_n \log(p_i) + \gamma_n \log(1 - t_{ni}) + \varepsilon_{ni}, \quad (17.9)$$

where p_i is the per-square-meter rental price of housing in municipality i , t_{ni} is the location- and income-specific progressive tax rate for household n in municipality i , and ε_{ni} is a household- and location-specific error term. Assuming that ε_{ni} follows an extreme value type I distribution leads to McFadden's (1974) conditional logit model. The parameters α_n and γ_n are modeled as linear functions of a vector x_n of observable household characteristics such as income and the number of children:

$$\alpha_n = \alpha_0 + \alpha_1 x_n \quad \text{and} \quad \gamma_n = \gamma_0 + \gamma_1 x_n. \quad (17.10)$$

This leads to interactions between location-specific and household-specific variables in the indirect utility function (17.9). Interaction effects and all other location- and household-specific variables are identified after introducing location-specific fixed effects

³⁹ See also [Kuminoff et al. \(2013\)](#) for a survey of the estimation of equilibrium sorting models.

θ_i absorbing all observed and unobserved location attributes that are equally important to all households:

$$V_{ni} = \theta_i + \alpha_1 \log(p_i) \cdot x_n + \gamma_0 \log(1 - t_{ni}) + \gamma_1 \log(1 - t_{ni}) \cdot x_n + \varepsilon_{ni}. \quad (17.11)$$

Note that the base effect γ_0 of housing prices and all other location-specific variables are not identified in Equation (17.11).

Schmidheiny (2006a) finds that local income tax rates are a highly significant and substantial determinant of household location choices. High-income households are more likely to move to low-tax municipalities. This is partly explained by the progressivity of the local tax rates. The results holds when controlling for observed social interactions including local average income and ethnic composition as explanatory variables.

17.4.2.2 Estimation of individual location choice in equilibrium

The empirical approach summarized in this section seeks to identify all parameters in the indirect utility function underlying the individual location choice including the base effects of location-specific variables such as housing prices. It also explicitly models how individual location choices affect location-specific characteristics in equilibrium, making it possible to perform counterfactual exercises.

The empirical strategy is borrowed from the empirical industrial organizations literature. Berry et al. (1995) introduced an econometric model to estimate the demand for cars based on their characteristics. Bayer et al. (2004) apply the Berry–Levinsohn–Pakes framework to the choice of neighborhoods j or individual housing units by households n .⁴⁰ The first step in Bayer et al. (2004) is the estimation of a multinomial choice model with the following indirect utility function:

$$V_{ni} = \alpha_n p_i + \beta_n x_i + \gamma_n z_i + \varepsilon_{ni}, \quad (17.12)$$

where p_i is the price of housing in neighborhood i , x_i are exogenous characteristics of the neighborhood j , and ε_{ni} is an individual- and location-specific error term. To allow for social interactions, they also include additional variables z_i with endogenous location characteristics such as average incomes and ethnic composition. In an empirical setting with local income taxes, the tax rates t_i would be another variable in z_i . The effects α_n , β_n , and γ_n are specific to the individual and are assumed to depend linearly on attributes x_n of individual n :

$$\alpha_n = \alpha_0 + \alpha_1 x_n, \quad \beta_n = \beta_0 + \beta_1 x_n, \quad \text{and} \quad \gamma_n = \gamma_0 + \gamma_1 x_n \quad (17.13)$$

⁴⁰ See also Sections 2.3 and 2.4 in the chapter by Holmes and Sieg in this handbook and Kuminoff et al. (2013) for more technical details and results.

The resulting indirect utility function is therefore

$$V_{ni} = \theta_i + \alpha_1 x_n \cdot p_i + \beta_1 x_n \cdot x_i + \gamma_1 x_n \cdot z_i + \varepsilon_{ni}, \quad (17.14)$$

where $\theta_i = \alpha_0 p_i + \beta_0 x_i + \gamma_0 z_i$. If ε_{ni} follows an extreme value type I distribution, this leads to estimating McFadden's (1974) conditional logit model with location fixed effects in the first step. The estimation in the first step typically implies the estimation of a large number of fixed effects θ_i , which can be numerically demanding in the maximum-likelihood estimation of Equation (17.14). Bayer et al. (2004) therefore propose a fixed-point algorithm that efficiently calculates the unique set of θ_i 's given the parameters α_1 , β_1 , and γ_1 for which the predicted shares of choosing neighborhood i in the sample $n = 1, \dots, N$ equal the observed shares. This is a property of the maximum-likelihood estimator of the conditional logit model, and the resulting θ_i 's are therefore maximum-likelihood estimates. The parameters α_1 , β_1 , and γ_1 are then estimated by maximizing a concentrated-likelihood function.⁴¹

In the second step of the estimation, the neighborhood fixed effects are regressed on the neighborhood characteristics:

$$\theta_i = \alpha p_i + \beta_0 x_i + \gamma_0 z_i + \eta_i. \quad (17.15)$$

Bayer et al. (2004) deal with the obvious endogeneity in the second step in the spirit of Berry et al. (1995) and use functions of exogenous characteristics in all other neighborhoods as instruments for local housing prices p_i . With the use of individual housing units as choice alternatives (see footnote 41), the endogeneity of social contextual variables z_i is addressed by restricting the analysis to a sample of houses near school attendance zone boundaries and including boundary fixed effects as in Black (1999). However, Bayer et al. (2004) do not deal with endogeneity in the first step.⁴²

The method of Bayer et al. (2004) makes it possible to perform counterfactual exercises. Changes in exogenous variables x_j will have a direct effect on the equilibrium location choices via β as well as an indirect effect through changing the endogenous location characteristics z_j . For example, the ethnic composition of a neighborhood is the aggregate of all individual location decisions in equilibrium. For counterfactual exercises, the effect of individual location choices on endogenous variables z_j has to be explicitly modeled and the new equilibrium has to be solved.

⁴¹ Bayer et al. (2004) estimate the first stage with individual housing units as choice alternatives. N individual households choose among N alternatives. With potentially hundreds of thousands of individuals and the same number of choice alternatives, the maximum-likelihood estimation is very expensive if not impossible. They therefore draw on a result obtained by McFadden (1978): the conditional logit model can be estimated on the basis of a random subset of choice alternatives for each individual. This choice set contains the actual choice plus a (small) random selection from the remaining alternatives. This estimation strategy depends on the independence of irrelevant alternatives (IIA) assumption.

⁴² Note that Berry et al. (1995) and the subsequent industrial organizations literature do not cause the endogenous variable p_j to interact with individual characteristics and do not include social interaction variables z_j .

Bayer et al. (2007) estimate the marginal willingness to pay for school quality. They find significant willingness to pay for increased school performance in the form of high housing prices but substantially smaller effects than previous estimates. Bayer et al. (2011) address the endogeneity of the location-specific variables z_i in Equation (17.14) by focusing on the relocation decision. They find that estimates based on a cross section of residents understate the willingness to pay for amenities such as air quality, but overstate the willingness to pay for living with one's own ethnic group.

To our knowledge, the Bayer et al. (2004) sorting strategy has not yet been used to study the impact of within-city tax differentials on population sorting.

17.4.2.3 Structural estimation

The theoretical models presented in Section 17.4.1.1 can be used for structural estimation.⁴³ Epple and Sieg (1999) take the theoretical model of Epple and Platt (1998), which introduced heterogeneity of households in terms of both income and tastes for public goods. This two-dimensional heterogeneity produces more realistic *partial* sorting in equilibrium: the residents of high-income municipalities are *on average* richer than those in low-income municipalities but the income distributions overlap. Different from Epple and Platt (1998), where the public good is a pure monetary transfer, public goods enter the utility function, leading to more realistic equilibria in which rich households sort into municipalities with high public good provision and high after-tax housing prices. Epple and Sieg (1999) fully parameterize the household's indirect utility function and the joint distribution of household heterogeneity (bivariate log-normal). In a first step, a subset of the structural parameters and a series of municipality fixed effects are estimated by matching the predicted income quartiles with the observed income quartiles across all municipalities in the city. In a second step, the municipality fixed effects are related to observed dimensions of public good provision (school quality, crime, parks, pollution, etc.) and per-unit housing prices using nonlinear least squares. The error term in this parameterization is unobserved public good provision, which is likely correlated with observed dimensions of public good provision and with housing prices. Instruments therefore need to be used for these variables in the second step. Note that different from the estimation in the Section 17.4.2.2, there is no idiosyncratic shock which lets households prefer different municipalities. Epple and Sieg (1999) estimate the model using data for the 92 cities and towns in the Boston metropolitan area in 1980. The estimated model fits the observed pattern of income sorting, housing prices, and public good provision remarkably well. The estimated model can be used to simulate the effect of property tax rates on the equilibrium location pattern.

⁴³ See also Sections 2.3 and 2.4 in the chapter by Holmes and Sieg in this handbook and Kuminoff et al. (2013) for a more technical discussion.

Epplé et al. (2001) also use the theoretical conditions of the majority voting equilibrium in the structural estimation. They find parameter estimates that are difficult to reconcile with the ones from the locational equilibrium in Epplé and Sieg (1999). Calabrese et al. (2006) structurally estimate the model by adding peer group effects and show that this eliminates the inconsistency between parameters based on the locational and the voting equilibrium. Epplé and Sieg (1999) assume that while households differ in their taste for the *level* of local public good provision, all households share the same valuation for the different *dimensions* of public good provision. Epplé et al. (2010) address this by allowing different types of households to have different valuation over the public good dimensions. We summarized the main results of this literature in Section 17.4.1.1.

The key finding of the three different empirical approaches is as follows:

Result 17.4.6. *There is empirical evidence that high-income households are attracted to low-income-tax states within countries and to low-income-tax municipalities within cities.*

All evidence to date on tax-induced population sorting at the local level is based on either microeconomic studies using observational data or structural estimation. Recent quasi-experimental evidence shows that very mobile and highly skilled workers are attracted to countries with low income tax rates.⁴⁴

17.5. TAXATION AND AGGLOMERATION ECONOMIES

So far in this chapter, while considering interactions among households, we have assumed that firms' location choices are mutually independent. This approach implies a presumption toward the spatial dispersion of economic activity, as density is deterred by the competition for inelastic housing. More broadly, to assume individual location choices to be independent is to ignore the central mechanism of urban economics: agglomeration forces. If agents generate positive externalities for other agents in the neighborhood, then activities will cluster in space and their sensitivity to taxes will not be the same as in the absence of such externalities. In this section, therefore, we aim to provide a summary of the theoretical literature and empirical literature on decentralized fiscal policy in the presence of agglomeration economies.

⁴⁴ Kleven et al. (2013) study the location choice of soccer players in Europe after the Bosman ruling by the European Court of Justice which lifted pre-existing restrictions on soccer player mobility. Kleven et al. (2014) analyze the effect of special tax breaks for high-income foreign workers in Denmark. Young and Varner (2011) study the effect of a substantial increase in the income tax rates on top earners in New Jersey. Closer to the focus of this survey, Agrawal and Hoyt (2013) use within-city state borders to identify tax effects on commuting times, and they find that city dwellers are indeed prepared to accept longer commutes in return for lower income tax rates.

17.5.1 Theory

The key implication of agglomeration economies for tax competition models is that economic activities, even if mobile in terms of the institutional setting, may be *de facto* immobile because in order to remain competitive firms need to locate at the industry cluster. Hence, policy makers can tax agglomerations without necessarily jeopardizing their tax base. This mechanism has been analyzed extensively in “new economic geography” models, featuring agglomeration equilibria in which a core region hosts the entire mobile sector that is subject to agglomeration forces while the periphery hosts some of the immobile industry only (Ludema and Wooton, 2000; Kind et al., 2000; Baldwin and Krugman, 2004; Krogstrup, 2008).⁴⁵ The key insight of this literature is that agglomeration forces make the world “lumpy”: when capital (or any other relevant production factor) is mobile and trade costs are sufficiently low, agglomeration forces lead to spatial concentrations of economic activity that cannot be dislodged by tax differentials, at least within certain bounds. In fact, agglomeration externalities create rents that can, in principle, be taxed by the jurisdiction that hosts the agglomeration. Moreover, decentralized fiscal policy can itself reinforce agglomeration tendencies when scale economies in the production of publicly provided goods make the locus of agglomeration even more attractive (Andersson and Forslid, 2003).⁴⁶ The core-periphery outcome, however, is quite extreme, particularly when considered at the scale of a city. It is therefore important to note that agglomeration economies need not be as stark as in the core-periphery case to reduce the intensity of tax competition. Borck and Pflüger (2006) show that local tax differentials can also be generated in models that produce stable equilibria with partial agglomeration, and where the mobile factor therefore does not derive an agglomeration rent.

Result 17.5.1. *Agglomeration economies can generate taxable rents and weaken the intensity of tax competition.*

While the mobility-reducing effect of agglomeration economies and the attendant attenuation of horizontal tax competition have been the most talked about policy insights generated by the new economic geography, the very same models in fact can generate the opposite result: knife-edge situations in which a very small tax differential can trigger large changes in the spatial distribution of the tax base. In those configurations, agglomeration economies in fact *add* to the sensitivity of firm location to tax differentials because one firm’s location choice can trigger further inflows and thus the formation of a new cluster. In such configurations, agglomeration economies exacerbate the intensity of tax competition (Baldwin et al., 2003, Result 15.8; Konrad and Kovenock, 2009). A similar result is found by Burbidge and Cuff (2005) and Fernandez (2005), who have studied tax competition in models featuring increasing returns to scale that are external to

⁴⁵ See Baldwin et al. (2003, chapters 15, 16) for an overview.

⁴⁶ The reverse mechanism, whereby decentralized fiscal policy favors economic dispersion, can be modeled as well, by considering the widely documented fact that public expenditure tends to be biased toward local suppliers (Brühlhart and Trionfetti, 2004).

firms, with firms operating under perfect competition. In these models, individual firm mobility is not constrained by agglomeration economies, and governments may compete even more vigorously to attract firms than in the standard tax competition model.

Result 17.5.2. *Potential agglomeration economies in spatially dispersed activities can imply large tax-base elasticities and thereby intensify tax competition.*

These results are essentially based on two-region models. In models featuring multiple regions, subtler differences emerge. [Hühnerbein and Seidel \(2010\)](#), using a standard new economic geography model, find that the core region might not be able to sustain higher tax rates in equilibrium if it is itself subdivided into competing jurisdictions. Similarly to the model of [Janeba and Osterloh \(2013\)](#), therefore, their model implies that tax competition puts particular pressure on central cities, which compete over mobile tax bases with other central cities as well as with their own hinterlands.

Such geography models hold particular promise for the analysis of tax policies within cities, given that production factors are highly mobile at that spatial scale and that agglomeration economies have been found to decay steeply over space ([Rosenthal and Strange, 2004](#)). If we focus on the scenario whereby locally stable clusters have already formed, such agglomeration forces could reduce race-to-the-bottom-type competitive pressures on local tax setting and thus make decentralized taxation efficient. It has furthermore been shown that decentralized tax setting can act as a mechanism of undoing inefficient spatial equilibria, where industry clusters are initially locked in a suboptimal location ([Borck et al., 2012](#)). Moreover, agglomeration economies may make decentralization more politically feasible, as they likely favor larger, central jurisdictions, thus giving central municipalities an advantage where in asymmetric models without agglomeration forces they generally are found as losing out from decentralization.

Result 17.5.3. *Agglomeration economies likely work to the advantage of central urban municipalities.*

The potential importance of agglomeration economies for urban public finance, therefore, is hard to overstate. However, firm-level agglomeration economies are not the only force that shapes intracity geographies. As we discussed in [Section 17.4.1.2](#), endogenous population sorting can lead to the geographically central municipality not being the economic center.⁴⁷

17.5.2 Empirical studies

An empirical assessment of the prediction for decentralized tax setting from recent theoretical work in economic geography boils down to three hierarchically nested questions

⁴⁷ Another interesting implication of agglomeration externalities is that they strengthen the theoretical case for some degree of intracity fiscal equalization ([Haughwout et al., 2002](#); [Riou, 2006](#); [Gagné and Riou, 2007](#); [Haughwout and Inman, 2009](#); [Wrede, 2014](#)).

(in the sense that the second and third of these questions are only relevant if the answer to the preceding question is positive):

1. Do firms internalize agglomeration economies in their location choices such that differences in tax burdens across locations become relatively less important (or more important, depending on the initial equilibrium)?
2. Do local governments realize that the mobility of their tax base is affected by agglomeration economies, and do they choose their tax rates accordingly?
3. Is the effect of agglomeration economies on local tax setting sufficiently strong to affect the equilibrium tax competition outcome significantly?

A number of empirical researchers have been looking for answers to these questions in recent years.

17.5.2.1 Do agglomeration economies make firms more or less sensitive to local taxation?

The first question boils down to testing the partial effect on a firm f 's location choice L_{fij} in location i and industry j of the local tax burden t_{fij} , of agglomeration effects a_{ij} , of a vector of other exogenous determinants x_{ij} , and of a random term ε_{fij} :

$$L_{fij} = g(t_{fij}, a_{ij}, t_{fij} \cdot a_{ij}, x_{ij}, \varepsilon_{fij}), \quad (17.16)$$

where L_{fij} equals 1 for the firm–location–industry cell corresponding to an actual location choice, and 0 for all other combinations of firm, location, and industry.⁴⁸ These models are typically estimated via conditional logit or Poisson count models, implying that g represents an exponential mean function (Schmidheiny and Brülhart, 2011; Brülhart and Schmidheiny, 2015). The key element of Equation (17.16) is the interaction term $t_{fij} \cdot a_{ij}$, which implies that the effects of taxation are not separable from the effects of agglomeration.⁴⁹

Devereux et al. (2007b) were the first to analyze the effect of both fiscal policy and agglomeration on location choices. They explore a variant of Equation (17.16) in which the measure of agglomeration is purely location specific, meaning that it can be written as a_i . Using data on British regional grants (their measure for t_{fij}), they find that, other things being equal, firms are more responsive to financial incentives in areas with pre-existing activity in the relevant industry. It may thus be cheaper to attract a new plant to an existing cluster than to a peripheral location. This is an important and evidently policy-relevant result, but not what the theory necessarily predicts. For an interior spatial

⁴⁸ Although most of the studies presented below include a time dimension on some or all of the variables, we abstract from it here in order to simplify the notation.

⁴⁹ A large body of empirical literature exists on variants of Equation (17.16) that do not feature the interaction term. See, for example, Hines (1999) for a survey, and de Mooij and Ederveen (2003) for a meta-analysis. Studies of the responsiveness of tax bases to tax rates at the local level include those of Buettner (2003) for Germany, Haughwout et al. (2004) for the United States, and Duranton et al. (2011) for the United Kingdom.

equilibrium with no relocation costs, expected profits at the locus of the agglomeration and at the periphery are equalized. Whether a given change in fiscal inducements is then more effective at attracting firms to a central location or to a peripheral location is indeterminate, as it depends on the functional form of the relationship between real returns and industry shares across locations. In the simulations reported by [Borck and Pflüger \(2006\)](#), for example, a given fiscal inducement will in fact attract a larger number of firms if offered at the peripheral location than if offered at the central location. Moreover, [Rohlin et al. \(2014\)](#) find that the deterrence effect of income taxes on firm location across US state borders is in fact stronger in denser areas.

[Brülhart et al. \(2012\)](#) explore this question by asking instead whether *industry*-level agglomeration economies reduce firms' sensitivity to local tax differentials. They estimate empirical location choice models for firm start-ups across Swiss municipalities. The distinctive feature of their model is an interaction term between local corporate tax rates and the [Ellison and Glaeser \(1997\)](#) index, a measure of industry-level agglomeration (a_j). Positive estimated coefficients on this interaction term imply that location choices of firms in more agglomerated sectors are less sensitive to tax differences across potential locations. By exploiting a setting in which municipal corporate taxes apply identically to firms across all sectors (such that taxes are not tailored to individual firms or sectors, allowing them to be written as t_i), and by instrumenting both tax rates and agglomeration measures, they seek to minimize potential endogeneity bias. They find that firm births, on average, react negatively to corporate tax burdens, but that the deterrent effect of taxes is weaker in sectors that are more spatially concentrated. Firms in sectors with an agglomeration intensity in the top quintile are less than half as responsive to differences in corporate tax burdens as firms in sectors with an agglomeration intensity in the bottom quintile. This finding supports the relevance of the theoretical prediction whereby agglomeration economies reduce the importance of tax differentials for firms' location choices.

[Jofre-Monseny and Solé-Ollé \(2012\)](#) expand on the approach of [Brülhart et al. \(2012\)](#) by estimating their regression model separately for Catalanian cities (local labor markets) featuring strong primacy of the central jurisdiction (defined as 40% or more of employment being concentrated in the largest municipality) and for more dispersed cities. Thereby, they seek evidence on the theoretical prediction whereby agglomeration forces can exacerbate the tax sensitivity of firm location if one is starting from a dispersed economic geography. Unlike [Brülhart et al. \(2012\)](#), they find the coefficient on the interaction term $t_i \cdot a_j$ to be significantly negative, and this particularly so in cities featuring strong primacy of the central municipality. Significant positive coefficient estimates on the interaction term are obtained only when the sample is limited to the central municipalities of the cities featuring strong jurisdictional primacy. These results appear to be in line with the theoretical prediction whereby, depending on the initial spatial configuration, agglomeration economies can strengthen or weaken firms' sensitivity to tax differentials.

Result 17.5.4. *The available evidence supports the prediction that agglomeration economies can make firms less sensitive to tax differentials across jurisdictions.*

This line of research leaves considerable scope for cross-validation and further elaboration. In particular, the dividing line between sensitivity-enhancing and sensitivity-reducing agglomeration economies could be fruitfully explored further, in particular by considering asymmetries not just in terms of aggregate jurisdiction size but also in terms of the initial-period spatial distribution of activity in the individual industries.

17.5.2.2 Do local-level tax policies take account of agglomeration economies?

The diagnosis that agglomeration economies exist and that they matter for firms' responses to tax differentials constitutes but the first step in a full evaluation of the prediction that agglomeration forces affect tax competition. The second question is whether policy makers recognize agglomeration forces and effectively seek to tax the associated rents or to compete all the more vigorously.

Most of the empirical literature in fact addresses this second question, taking the offsetting effect of agglomeration economies on firms' sensitivity to tax differentials as a given. These studies estimate models of the following type:

$$t_{ij} = h(a_{ij}, x_{ij}, \varepsilon_{fij}), \quad (17.17)$$

where x_{ij} now stands for exogenous determinants of local tax rates other than agglomeration economies a_{ij} , and h typically represents a linearly additive function.

The results of this literature are easily summarized: all of the existing studies conclude that observed tax rates are higher in places that are identified by researchers as hosting an agglomeration. This is particularly pronounced for the early studies: Buettner (2001) finds that more populous German municipalities set higher local business tax rates, and Charlot and Paty (2007) observe that French municipalities with greater market potential set higher business tax rates. This means that they find large and statistically significant coefficients on location-specific agglomeration measures, a_i . The estimates of Charlot and Paty (2007), for instance, imply that a 10% increase in market access increases the business tax rate by 1.3% on average.

Such analyses have to contend with formidable empirical challenges. One issue is the potential for reverse causation, whereby t affects A rather than the other way around—a theoretically well established link (e.g., Andersson and Forslid, 2003). More recent studies have sought to allay this problem by using as an instrument for location-specific agglomeration measures a_i agglomeration measured at a date prior to the introduction of the left-hand-side variable t_i (Jofre-Monseny, 2013; Koh et al., 2013; Luthi and Schmidheiny, 2014). While these approaches go a long way toward allaying the reverse-causation concern, one cannot rule out that some related tax instrument existed in the past and played a part in determining agglomeration patterns.

A probably even greater empirical challenge arises from local revenue needs as a confounder of agglomeration. Larger, denser, and more central locations invariably correspond to more urban places, and central cities are typically associated not only with agglomeration economies but also with stronger demand for publicly provided goods. Researchers typically try to control for as many observables as possible, by including vectors of sociodemographic characteristics among the location-level controls x_i . Yet, a lot is asked of these controls if they are to filter out differences in demands for publicly provided goods completely.

Jofre-Monseny (2013) and Koh et al. (2013) have addressed this issue by considering not just aggregate density (a_i) but also location-industry-level agglomeration measures (a_{ij}), thus adding an industry dimension that is in principle orthogonal to the problematic location dimension. Both studies find measures of a_{ij} to be associated with significantly higher average local-level tax rates as well.

Luthi and Schmidheiny (2014) in addition distinguish differentials *across* cities from differentials *within* cities (defined as Swiss metropolitan areas). They observe that between cities, both jurisdictional size and centrality—two alternative measures of a_i —are associated with higher tax rates. This is consistent with asymmetric tax competition models as well as with core-periphery models. Within cities, however, only jurisdictional size appears to matter, whereas proximity to the center (conditionally on size) is not significantly correlated with observed tax rates. Importantly, given the focus of this chapter, the study authors interpret their finding as evidence that the standard asymmetric tax competition mechanism is at play both within and among cities, but that the agglomeration mechanism seems to matter only for tax differentials across cities. To our knowledge, this is the only study so far to have distinguished intraurban from interurban determinants of local tax setting while considering agglomeration economies. This seems to be a promising area for further research.

Another approach to addressing potential omitted-variable bias due to heterogeneous revenue needs is adopted by Brühlhart and Simpson (2015). They take advantage of the fact that British regional development subsidies, interpreted as inverse taxes, can be varied across firms, thus yielding a dependent variable that can be denoted as t_{fij} . With such a regressand, identification can come from the industry dimension instead of the location dimension. Brühlhart and Simpson (2015) test whether subsidies requested by applicant firms offered by the government take account of firms' differential spatial mobility according to the extent of industry localization measured through the Ellison and Glaeser (1997) index. They find evidence of firms internalizing agglomeration economies in their applications and of government agencies reflecting this in the generosity of their subsidy offers.

However, they also observe that local government agencies structure their offers so as to try to preserve existing employment in more agglomerated industries at the locus of agglomeration. Such behavior corresponds better to theories of policy capture by

dominant incumbent industries than to geography models in which disinterested local governments shape their tax policy in order to account for agglomeration effects. Put simply, while agglomerations in principle are taxable, they might leverage their weight in local economies to obtain favorable tax treatment. This is consistent with political-economy theories according to which policy capture by vested interest is stronger at the local level than at the national level (Bardhan and Mookherjee, 2000; Redoano, 2010).

Result 17.5.5. *Larger and more central municipalities are generally found to apply higher tax rates. The extent to which this reflects taxable agglomeration rents remains moot.*

17.5.2.3 Do agglomeration economies affect the equilibrium tax competition outcome significantly?

It would seem fair to summarize the relevant literature as yielding a cautious “yes” to the first two questions: firms in agglomerated sectors trade off higher taxes for greater proximity to other firms, and local governments seem to recognize this to some extent and set their taxes accordingly. Does this mean that agglomeration forces significantly counter-balance race-to-the bottom forces of horizontal tax competition within and between urban areas? And how relevant are agglomeration forces for personal rather than corporate tax bases (i.e., local sorting effects; see Section 17.4)? These questions have so far remained unaddressed. A rigorous treatment would likely require structural modeling allowing counterfactual simulations for different agglomeration intensities.

17.6. CONCLUDING REMARKS

As shown in Section 17.2, the typical OECD city is divided into 74 municipalities of on average 20,000 inhabitants when we define a “city” as a functional urban area of at least 500,000 residents. This population average masks huge size variations: the typical central municipality accounts for fully 40% of the city’s population and is thus some 50 times bigger than its average surrounding urban municipality. One key task of these urban jurisdictions is to raise tax revenue amounting to 10% of consolidated (local, regional, and national) taxation. The representative urban municipality raises 43% of its revenue from property taxes, 21% from taxes on personal income, and 21% on taxes on the consumption of goods and services. The dominance of local property taxes is a feature mainly of English-speaking countries—in countries with the highest levels of local tax autonomy, personal income taxes tend to dominate even at the municipal level.

These stylized facts make it abundantly clear that models of tax competition, although originally framed in an international setting, hold considerable relevance also for tax setting within urban areas, where multiple horizontally and vertically nested jurisdictions of very different sizes compete at close quarters for a range of mobile tax bases. We have

shown that different plausible models have sharply differing implications for the relationship between municipal population and tax rates, and that quintessentially urban features need to be better incorporated into these models.

Intraurban tax competition is different from intercity and international tax competition in that all tax bases are highly mobile within cities, including private households. This means that taxation within cities shapes and is shaped by residential sorting. The decentralized provision and financing of public goods within cities allows rich households to sort into rich municipalities with high public good provision, low tax rates, and high housing prices that prevent low-income households from following. Such Tiebout-type sorting is potentially efficient, as different (income) groups consume public goods tailored to their preferences. However, calibrated theory models show that such welfare gains turn out to be elusive in the case of municipal property taxation.

While firms and households are mobile within cities and to some extent also between cities within a country, this mobility can be constrained by agglomeration forces. This phenomenon has been subject to particular scientific scrutiny in recent years, owing to the prediction of “new economic geography” models that agglomeration forces can make firms *de facto* immobile and thereby generate taxable location rents. We reviewed this literature in [Section 17.5](#) and found that, while agglomeration forces could in theory both intensify and attenuate tax competition, the evidence points toward an attenuating effect, as agglomeration forces are found to reduce firms’ sensitivity to local tax differentials. Whether this mechanism is of first-order importance in determining local tax rates, however, remains uncertain.

The literature on intraurban tax setting is still patchy, and many articles we discussed in this survey, while relevant to the issue, are not intentionally aimed at shedding light on this particular problem. Moreover, most of the literature we have covered focuses on positive theoretical predictions and their support in the data. Robust welfare-relevant results, however, are scarce, as is empirical work looking specifically at policy making by intraurban jurisdictions. Given the global trend toward urbanization and, in many countries, fiscal decentralization, this surely offers a fruitful area for further research.

APPENDIX

The data on jurisdictional fragmentation of consistently defined urban areas are from the OECD data Web site at <http://stats.oecd.org> under the theme “Regions and Cities” and the subtheme “Metropolitan areas” (DOI 10.1787/region-data-en). We use the variables “total population metro area (persons),” “local governments (count),” and “territorial fragmentation.” We exported data for 2012 which reports population figures for 2012 and the number of local governments from various years. See [OECD \(2013a, p. 174\)](#) and [Table A.1](#) for reported years by country.

The data on local fiscal decentralization for all countries except for the United States are from the IMF at <http://elibrary-data.imf.org>. Note that GFS data are accessible only with a subscription. We downloaded the data using the IMF query builder through the following steps: “Sign in” with user name and password. “Query within a dataset: Government Finance Statistics (GFS).” Choose “Time” 1960–2012 (we downloaded 10-year intervals to limit the size of the individual datasets). Choose “Unit,” tick “National currency” and “Euros,” choose “Concept,” expand “2001 GFS,” expand “Cash,” mark “Local Government,” click “Select Branch.” Repeat with “Noncash” and with “General Government” for both “Cash” and “Noncash” data. This choice results in the selection of 1666 of 7548 items. Choose “Country,” click “Select All.” Export data.

The index LTS in Equation (17.1) is calculated from the GFS variable GLRT_G01_AC for the numerator “local government tax revenue,” and GGRT_G01_AC for the denominator “general government tax revenue.” For countries which do not report noncash budgetary information, we used GLRT_G01_CA and GGRT_G01_CA, respectively. We used the most current observation for which local data are reported. We used only countries for which the local data do not include sub-central government units such as states and regions, which are clearly larger than urban areas. See Table A.3 for the list of included countries and years and Table A.4 for the list of excluded countries. The composition of local taxes is taken from the variables GLRTII_G01_AC for personal income tax, GLRTIC_G01_AC for corporate income tax, GLRTP_G01 for property tax, and GLRTGS_G01 for tax on consumption, and the residual $GLRT\_G01\_AC - GLRTII\_G01\_AC - GLRTIC\_G01\_AC - GLRTP\_G01\_AC - GLRTGS\_G01\_AC$ for other tax sources. For countries that do not report noncash budgetary information, we use the _CA versions of the variables. The index ALTS in Equation (17.3) uses the share of local tax revenue in tax groups (a) plus (b1) plus (b2) plus (c) from Blöchliger and Rabesona (2009, p. 5, Table A.2).

The GFS data offer no information on local fiscal decentralization in the United States after 2001. We therefore used data from the Historical Finance Data Base (IndFin) provided by the US Census Bureau. This dataset reports time series of financial variables from 1967 to 2011 on an annual basis. IndFin is not publicly accessible, but access can be requested by e-mail: govs.census.management@census.gov. We used data for 2007, the most recent year with data on the universe of local units. We used the variable `totaltaxes` for total tax revenue, `individualincometax` for personal income tax revenue, `corpnetincometax` for corporate income tax revenue, `propertytax` for property tax revenue, and `totsalesgrrectax` for consumption tax revenue. Revenue from other tax sources was calculated as the residual between total taxes and the four components. Tax revenue of all local governments was calculated as the aggregate of all revenue from government units for which the variable `typecode` takes values 2 (municipality), 3 (township), 4 (special district), or 5 (school district, independent only). Note that IndFin data report annual cash flows only and therefore correspond to the

Table A.1 Local government units in OECD/EC data

Country	ISO country code	Year	Source	Local governments reported by OECD
Austria	AUT	2001	Eurostat	Gemeinden (LAU2)
Belgium	BEL	2001	Eurostat	Gemeenten/communes (LAU2)
Canada	CAN	2006	Statcan	Census subdivisions (towns, villages, etc.) (CSD)
Chile	CHL	2002	INE	Chile, comunas
Czech Republic	CZE	2001	Eurostat	Obce (LAU2)
Denmark	DNK	2001	Eurostat	Sogne (LAU2)
Estonia	EST	2000	Eurostat	Vald, linn (LAU2)
Finland	FIN	2000	Eurostat	Kunnat/kommuner (LAU2)
France	FRA	1999	Eurostat	Communes (LAU2)
Germany	DEU	2001	Eurostat	Gemeinden (LAU2)
Greece	GRC	2001	Eurostat	Demotiko diamerisma/koinotiko diamerisma (LAU2)
Hungary	HUN	2001	Eurostat	Települek (LAU2)
Ireland	IRL	2001	Eurostat	Local governments (LAU1)
Italy	ITA	2001	Eurostat	Comuni (LAU2)
Japan	JPN	2006	NLFTP	Shi (city), machi or cho (town), and mura or son (village)
Korea	KOR	2009	KOSIS	Eup, myeon, dong
Luxembourg	LUX	2001	Eurostat	Communes (LAU2)
Mexico	MEX	2010	INEGI	Municipios
Netherlands	NLD	2001	Eurostat	Gemeenten (LAU2)
Norway	NOR	2001	Eurostat	Municipalities (LAU2)
Poland	POL	2002	Eurostat	Gminy (LAU2)
Portugal	PRT	2001	Eurostat	Freguesias (LAU2)
Slovak Republic	SVK	2001	Eurostat	Obce (LAU2)
Slovenia	SVN	2002	Eurostat	Obeine (LAU2)
Spain	ESP	2001	Eurostat	Municipios (LAU2)
Sweden	SWE	2000	Eurostat	Kommuner (LAU2)
Switzerland	CHE	2000	Eurostat	Municipalities (LAU2)
United Kingdom	GBR	2001	ONS	County councils
United States	USA	2000	US Census	Municipalities or townships

Year in which the local government units are counted. The local governments used in this report were identified on the basis of the following criteria: Have only one level of local government per country, notably the lowest tier (even if more than one level of government may have relevant responsibilities over the same territory). Identify only general-purpose local governments, excluding the specific function governments (for example, school district, health agencies, etc.). For the United Kingdom, for those areas where the county councils were abolished, the local authority (either a metropolitan district council or a unitary district council) is used. For London, the borough councils are used. For the United States, in the geographic areas where municipalities or townships do not represent a general purpose government, the county governments were considered.

Source: (OECD, 2013a, p. 174).

Table A.2 Jurisdictional fragmentation across OECD countries

Country	ISO country code	Functional urban areas Population			Number of local governments per area			Local governments per 100,000 inhabitants		
		No.	Average	Max.	Average	Min.	Max.	Average	Min.	Max.
Austria	AUT	3	1,323,321	2,737,753	209	140	313	20.8	11.4	28.0
Belgium	BEL	4	1,230,263	2,536,106	50	24	99	4.3	3.0	6.1
Canada	CAN	9	2,181,109	6,671,162	60	6	191	3.5	0.8	6.9
Chile	CHL	3	2,803,954	6,531,598	21	6	47	0.8	0.6	1.0
Czech Republic	CZE	3	1,024,677	1,868,631	249	67	435	24.4	11.9	38.1
Denmark	DNK	1	2,007,352	2,007,352	57	57	57	2.8	2.8	2.8
Estonia	EST	1	530,640	530,640	28	28	28	5.3	5.3	5.3
Finland	FIN	1	1,476,662	1,476,662	22	22	22	1.5	1.5	1.5
France	FRA	15	1,706,750	11,862,466	280	35	1375	21.4	6.3	49.1
Germany	DEU	24	1,321,825	4,386,551	74	3	308	5.2	0.4	11.0
Greece	GRC	2	2,256,708	3,547,773	61	28	94	2.8	2.7	2.9
Hungary	HUN	1	2,862,326	2,862,326	183	183	183	6.4	6.4	6.4
Ireland	IRL	1	1,735,182	1,735,182	7	7	7	0.4	0.4	0.4
Italy	ITA	11	1,672,074	4,109,109	70	14	252	3.9	2.6	6.1
Japan	JPN	36	2,426,972	35,441,287	22	3	235	1.2	0.3	2.5
Korea	KOR	10	3,660,358	23,496,373	172	27	965	5.4	4.1	6.4
Mexico	MEX	33	1,807,044	19,802,161	9	1	55	0.7	0.1	6.0
Netherlands	NLD	5	1,244,345	2,406,043	27	11	57	2.1	1.2	2.7
Norway	NOR	1	1,261,977	1,261,977	30	30	30	2.4	2.4	2.4
Poland	POL	8	1,433,687	3,008,921	41	17	101	2.9	1.8	4.9
Portugal	PRT	2	2,073,419	2,840,065	193	150	235	9.9	8.3	11.5
Slovakia	SVK	1	722,106	722,106	136	136	136	18.8	18.8	18.8
Slovenia	SVN	1	576,370	576,370	28	28	28	4.9	4.9	4.9
Spain	ESP	8	2,126,111	6,779,528	94	16	272	6.0	2.0	24.7
Sweden	SWE	3	1,181,950	1,991,310	18	12	28	1.6	1.4	2.1
Switzerland	CHE	3	935,770	1,226,332	167	140	193	19.0	11.4	23.9
United Kingdom	GBR	15	1,721,399	12,090,254	8	3	47	0.5	0.3	0.8
United States	USA	70	2,400,635	17,378,937	82	2	540	4.7	0.2	21.5

Source: Own calculations based on OECD (Regional Statistics). Population data from 2012, number of local governments from various years (see [Table A.1](#)).

Table A.3 Local government units in GFS/IndFin data, included countries

Country	ISO country code	Year	Local government units reported by IMF
Argentina	ARG	2004	1617 municipalities
Armenia	ARM	2012	900 marzes or communities
Australia	AUS	2012	900 cities, district councils, municipalities, shires, and towns
Austria	AUT	2011	2358 municipalities (excluding Vienna), municipal associations (education services), Vienna
Belgium	BEL	2011	589 communes
Bolivia	BOL	2007	9 municipalities of departmental capitals and numerous other municipalities
Bosnia–Herzegovina	BIH	2012	4 cities and 140 municipalities
Brazil	BRA	2011	5564 local governments
Bulgaria	BGR	2011	264 municipalities
Canada	CAN	2012	Municipal governments
Cape Verde	CPV	2009	22 municipalities, 3 municipal associations, and 15 water supply and sanitation agencies
Chile	CHL	2012	Municipalities and municipal mutual fund.
Colombia	COL	2011	1108 municipalities, including the municipality of Bogotá
Costa Rica	CRI	2007	81 municipalities
El Salvador	SLV	2011	262 municipalities
Estonia	EST	2011	39 city councils and 202 municipalities
Finland	FIN	2011	432 municipalities
Germany	DEU	2011	15,000 municipalities and municipal associations
Greece	GRC	2011	1033 communities and municipalities
Honduras	HND	2012	298 municipalities
Iceland	ISL	2011	17 municipalities, including public nursery and primary schools, and old persons' residential institutions
Iran	IRN	2009	1000 municipalities
Jamaica	JAM	2005	Kingston and St. Andrew Corp., Municipal Services Commission, Parish Council Services Commission, and 13 parish councils
Jordan	JOR	2011	Greater Amman municipality, 172 municipalities, and 350 village councils
Latvia	LVA	2012	109 amalgamated municipalities and 7 major towns
Lithuania	LTU	2012	60 local governments and nonprofit institutions (nursing homes, pre-primary, primary, and secondary schools, etc.), which are controlled and mainly financed by local governments. Municipal enterprise Vilniaus Miesto Bustas
Luxembourg	LUX	2011	116 communal administrations and municipalities

Continued

Table A.3 Local government units in GFS/IndFin data, included countries—cont'd

Country	ISO country code	Year	Local government units reported by IMF
Malaysia	MYS	2001	2 agencies with the functions of a local government, 12 city councils, 38 municipal councils, and 96 district councils. States reported separately
Malta	MLT	2011	68 local councils
New Zealand	NZL	2011	86 local government units
Paraguay	PRY	2012	Capital and 239 municipalities
Peru	PER	2012	7 decentralized agencies, 194 provincial councils, and 1836 district councils. Regions reported separately
Russia	RUS	2012	24,255 local governments
Slovenia	SVN	2011	210 municipalities
South Africa	ZAF	2011	6 metropolitan municipalities, 46 district municipalities, and 231 local municipalities
Spain	ESP	2012	9000 municipalities and other local authorities
Swaziland	SWZ	2003	2 city councils, 3 town boards, and 3 town councils
Switzerland	CHE	2010	2600 communes
United Kingdom	GBR	2012	540 local councils and local government units
United States	USA	2007	19,484 cities, 16,475 townships, 35,574 special districts, and 13,742 school districts

Year means latest observation with revenue data at the local level.

Source: IMF (Government Finance Statistics Yearbook, various years).

variables with suffix _CA in the GFS data. The tax revenue of the general government in 2007 was taken from p. 42 of the Financial Statements of the United States Government (downloaded from <https://www.fms.treas.gov/fr/07frusg/07stmt.pdf>). We calculated the index ALTS in Equation (17.3) for the United States using the share of local tax revenue in tax groups (a) plus (b) plus (c) from Stegarescu (2006, p. 32, Table 2.2).

To assess whether the IndFin and GFS accounting frameworks are comparable, we calculated the decentralization indices for the United States in 1987 and 1992 from both data sources. Both datasets report data based on cash flows. In 1987, the index LTS equals 12.8% in the IndFin data and 16.8% in the GFS data. In 1992, LTS is 14.2% in the IndFin data and 18.9% in the GFS data. While this systematic underestimation of decentralization in the IndFin data is substantial, it is not different by orders of magnitude and does not substantially change the comparison of tax decentralization across countries.

The OECD data on jurisdictional fragmentation at <http://stats.oecd.org> do not contain population figures for individual municipalities within cities (urban areas). We therefore additionally used a dataset provided by the EC (Urban Audit) which lists names and 2006 population figures for all individual municipalities within European

Table A.4 Local government units in GFS/IndFin data, excluded countries

Country	ISO country code	Year	Local government units reported by IMF
Afghanistan	AFG	2011	Partial information was provided. Municipalities. No state level reported
Azerbaijan	AZE	2012	51 districts (rayons) consisting of 5 cities and 1494 municipalities; the city of Baku, which consists of 52 municipalities; 171 Nakhichevan Autonomous Republic municipalities grouped into 7 districts and the city of Nakhichevan
Belarus	BLR	2012	1 city of Minsk, 12 oblast cities, 6 oblast/provincial, 118 rayon/districts, rayon 14 cities, 1289 rural, and 64 settlement/townships budgets
China	CHN	2011	656 cities, 2487 counties, 31 provinces (excluding Taiwan, Hong Kong, and Macao and including Beijing, Shanghai, Chongqing, and Tianjin), 333 subprovincial administrative regions, 44,067 townships, and 678,589 villages
Republic of the Congo	COG	2005	Brazzaville and 5 other municipalities, 11 departments
Croatia	HRV	2011	1 city (Zagreb), 20 counties, 126 towns, and 429 municipalities
Cyprus	CYP	2011	6 districts, 33 municipalities, and 298 village authorities
Czech Republic	CZE	2011	8 regional committees of cohesion, 14 regions, and 6300 municipalities
Denmark	DNK	2011	5 regions, 98 municipalities, regional and municipal agencies
France	FRA	2012	100 departments, 26 regions, and 36,000 communes
Georgia	GEO	2012	Adjara Autonomous Republic, Abkhazia Autonomous Republic, and 62 administrative districts, towns, and cities
Hungary	HUN	2012	19 county governments, 3200 municipalities and local minority governments
Ireland	IRL	2011	5 borough councils, 5 city councils, 29 county councils, and 75 town councils, as well as 2 regional assemblies and 8 regional authorities
Israel	ISR	2012	260 local government units (local councils, municipalities, regional councils)
Italy	ITA	2011	2 autonomous provinces, 20 regions, 104 provinces, 311 municipalities' unions, and 8101 municipalities
Japan	JPN	2011	1800 local public entities
Kazakhstan	KAZ	2011	2 cities (Almaty and Astana) and 14 oblasts (province) bodies.

Continued

Table A.4 Local government units in GFS/IndFin data, excluded countries—cont'd

Country	ISO country code	Year	Local government units reported by IMF
Korea	KOR	2011	6 metropolitan cities, 8 provinces, 69 autonomous districts, 73 cities, 86 counties, 228 basic local government units, special self-governing province Jeju, and special metropolitan city Seoul
Lesotho	LSO	2008	Maseru Municipal Council and district councils
Mauritius	MUS	2012	4 district councils and 5 municipal councils
Mexico	MEX	1998	Federal district, 31 state governments, and 2418 municipal governments
Moldova	MDA	2012	8393 local government budgetary organizations, 2 municipalities, and 32 rayons
Mongolia	MNG	2012	9 districts (ulaanbaatar), 21 provinces (aimags), and 331 districts (soums)
Morocco	MAR	2011	41 arrondissements, 61 prefectures and provinces, 16 regions, 1298 rural communes, 132 trade unions, and 200 urban communes
Netherlands	NLD	2012	316 communal arrangements, 418 municipalities, 12 provinces, and 26 public water boards
Norway	NOR	2012	18 counties and 430 municipalities
Poland	POL	2011	16 districts, 372 counties, and 2478 communes
Portugal	PRT	2012	Regional governments of the Azores and Madeira, 18 districts, and 308 municipalities
Romania	ROU	2011	1 city (Bucharest), 41 counties, 103 municipalities, 211 towns, and 2850 communes
Serbia	SRB	2011	Autonomous Province Vojvodina, cities, and municipalities
Slovakia	SVK	2011	8 regions and 2900 municipalities and other units
Sweden	SWE	2012	20 county councils, 186 municipal associations, and 290 municipalities
Tajikistan	TJK	2004	3 provinces, 15 cities, and 54 districts
Thailand	THA	2002	75 changwat administrative organizations, 1129 municipalities, 6745 district administrative organizations, Bangkok metropolitan administration, and Pattaya City.
Tunisia	TUN	2011	24 government councils and 264 municipalities
Ukraine	UKR	2011	1 republic (Crimea), 2 cities (Kiev and Sevastopol), 24 oblasts, 176 municipalities, and 488 districts
West Bank and Gaza	PSE	2010	428 localities/municipalities and 16 main governorates

Year means latest observation with revenue data at the local level.

Source: IMF (Government Finance Statistics Yearbook, various years).

Table A.5 Decentralization of local (municipal) revenue

Country	ISO country code	Year	LTS						
			Total	Personal income	Corporate income	Property	Consumption	Other	ALTS
OECD countries									
Australia	AUS	2012	3.4	0.0	0.0	3.4	0.0	0.0	3.4
Austria	AUT	2011	18.0	4.9	1.2	1.7	6.9	3.3	1.5
Belgium	BEL	2011	7.9	2.8	0.0	4.3	0.7	0.0	7.4
Canada	CAN	2012	11.3	0.0	0.0	11.0	0.3	0.0	11.0
Chile	CHL	2012	7.0	0.0	0.0	2.9	4.1	0.0	
Estonia	EST	2011	21.4	19.0	0.0	1.6	0.8	0.0	
Finland	FIN	2011	33.0	28.1	2.8	2.1	0.0	0.0	30.3
Germany	DEU	2011	13.0	5.0	5.4	1.7	0.9	0.0	7.8
Greece	GRC	2011	1.1	0.0	0.0	0.8	0.3	0.0	0.6
Iceland	ISL	2011	30.1	24.3	0.0	5.3	0.4	0.0	28.8
Luxembourg	LUX	2011	6.7	0.0	6.2	0.4	0.1	0.0	6.6
New Zealand	NZL	2011	7.4	0.0	0.0	6.9	0.5	0.0	0.0
Slovenia	SVN	2011	18.1	14.6	0.0	2.7	0.8	0.0	
Spain	ESP	2012	15.2	2.3	0.5	6.3	6.0	0.0	11.0
Switzerland	CHE	2010	20.1	13.7	2.5	2.9	0.5	0.5	20.1
United Kingdom	GBR	2012	6.0	0.0	0.0	6.0	0.0	0.0	6.0
United States	USA	2007	15.3	0.8	0.3	11.1	2.3	0.9	15.3
Non-OECD countries									
Argentina	ARG	2004	0.2	0.0	0.0	0.0	0.0	0.2	
Armenia	ARM	2012	2.6	0.0	0.0	2.1	0.5	0.0	
Bolivia	BOL	2007	21.1	0.0	2.3	5.1	12.6	1.0	
Bosnia–Herzegovina	BIH	2012	11.1	2.0	0.0	1.2	7.8	0.1	
Brazil	BRA	2011	7.2	0.6	0.0	2.2	3.1	1.2	
Bulgaria	BGR	2011	7.2	0.1	0.0	7.0	0.1	0.0	
Cape Verde	CPV	2009	3.3	0.0	0.0	2.3	0.9	0.0	
Colombia	COL	2011	11.1	0.0	0.0	4.1	7.0	0.0	
Costa Rica	CRI	2007	3.4	0.0	0.0	1.1	2.1	0.2	
El Salvador	SLV	2011	5.8	0.0	0.0	0.0	5.8	0.0	

Continued

Table A.5 Decentralization of local (municipal) revenue—cont'd

LTS									
Country	ISO country code	Year	Total	Personal income	Corporate income	Property	Consumption	Other	ALTS
Honduras	HND	2012	4.0	0.0	0.0	4.0	0.0	0.0	
Iran	IRN	2009	15.6	0.0	0.0	0.0	3.6	12.0	
Jamaica	JAM	2005	1.3	0.0	0.0	0.8	0.5	0.0	
Jordan	JOR	2011	2.5						
Latvia	LVA	2012	27.8	23.5	0.0	4.0	0.2	0.2	
Lithuania	LTU	2012	16.3	13.5	0.0	1.8	0.4	0.6	
Malaysia	MYS	2001	3.3						
Malta	MLT	2011	0.0	0.0	0.0	0.0	0.0	0.0	
Paraguay	PRY	2012	5.1	0.0	0.0	2.5	1.8	0.8	
Peru	PER	2012	2.3	0.0	0.0	1.1	0.2	1.0	
Russia	RUS	2012	6.1	4.3	0.1	1.0	0.1	0.7	
South Africa	ZAF	2011	4.3	0.0	0.0	4.2	0.1	0.0	
Swaziland	SWZ	2003	1.8	0.0	0.0	1.7	0.1	0.0	
Average			10.0	4.2	0.6	3.1	1.9	0.6	10.7
OECD			13.1	6.4	1.1	4.0	1.4	0.3	10.7
Non-OECD			7.4	2.2	0.1	2.3	2.3	0.9	
Minimum			0.0	0.0	0.0	0.0	0.0	0.0	0.0
Maximum			33.0	28.1	6.2	11.1	12.6	12.0	30.3

Indices in percentage points. Year means latest observation with data at the local (municipal) level.

Source: Own calculations based on data by the IMF (GFS) and US Census (IndFin).

Table A.6 Taxing autonomy of local (municipal) governments

Country	ISO country code	Share of local tax revenue in class							
		(a)	(b)	(c)	(d.1)	(d.2)	(d.3)	(d.4)	(e)
Australia	AUS	100.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Austria	AUT	2.6	5.5	0.0	0.0	0.0	65.3	0.0	20.7
Belgium	BEL	8.4	85.7	0.0	0.0	0.0	0.0	0.0	5.8
Canada	CAN	1.8	95.6	0.0	0.0	0.0	0.0	0.0	1.6
Finland	FIN	0.0	91.8	0.0	0.0	0.0	0.0	8.1	0.2
Germany	DEU	0.0	59.7	0.0	0.0	39.4	0.0	0.0	0.0
Greece	GRC	0.0	53.9	0.0	0.0	0.0	0.0	0.0	46.1
Iceland	ISL	0.0	95.9	0.0	0.0	0.0	0.0	0.0	0.0
Luxembourg	LUX	98.5	0.2	0.0	0.0	0.0	0.0	0.0	1.1
New Zealand	NZL	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
Spain	ESP	22.6	49.5	0.0	0.0	17.4	0.0	0.0	5.3
Switzerland	CHE	3.0	97.0	0.0	0.0	0.0	0.0	0.0	0.0
United Kingdom	GBR	0.0	100.0	0.0	0.0	0.0	0.0	0.0	0.0
United States	USA	71.4	28.6	0.0	0.0	0.0	0.0	0.0	0.0

Source: Blöchliger and Rabesona (2009) and Stegarescu (2006). Shares in percentage points.

cities. This dataset is not publicly available and was kindly provided to us by Lewis Dijkstra, Deputy Head of the Analysis Unit in the EC Directorate-General for Regional Policy. These data are based on the joint EC/OECD definition of cities. However, there are unfortunately substantial differences in the public OECD data used in [Section 17.2.1](#) and the EC data used in [Section 17.2.4](#). See [footnote 13](#), for example. We only used data on OECD countries as in [Section 17.2.1](#). We dropped the United Kingdom, Ireland, Denmark, Slovakia, and Portugal because the EC reports smaller units than the municipalities counted in the OECD data as local governments. We also dropped the Czech Republic, because the number of local governments differs considerably from the numbers in the OECD data. We used the same sample of cities as in the OECD data in [Section 17.2.1](#)—that is, we included some cities with a population below 500,000 inhabitants in 2006 but above that level in 2009; we also excluded some cities with a population above 500,000 inhabitants in the EC data but not included in the OECD data. We used the 2006 population figure for the largest and the second largest municipality for each city as well as the 2006 total population. The population share of the largest municipality was calculated as the population of the largest municipality divided by the total population of the city (urban area). The average municipality size in an urban area was calculated as the 2006 population in the EC data divided by the number of local jurisdictions in the EC data. The Zipf prediction in [footnote 15](#) is based on the 2006 population of the largest municipality in the EC data and the 2012 total population of the urban area in the OECD data.

Table A.7 Asymmetries across local governments within urban areas

Country	ISO country code	Population share of the largest			Largest vs. average			Largest vs. second largest			Largest vs. Zipf prediction		
		Mean	Min	Max	Mean	Min	Max	Mean	Min	Max	Mean	Min	Max
Austria	AUT	46	31	64	105.2	43.5	199.0	23.9	6.6	43.3	2.7	1.7	4.0
Belgium	BEL	30	7	46	11.0	7.3	14.6	5.6	1.7	10.4	1.2	0.4	1.8
Estonia	EST	73	73	73	17.6	17.6	17.6	29.0	29.0	29.0	2.8	2.8	2.8
Finland	FIN	43	43	43	6.1	6.1	6.1	2.4	2.4	2.4	1.4	1.4	1.4
France	FRA	31	15	50	74.3	10.5	268.1	7.0	2.2	19.1	1.8	1.0	2.7
Germany	DEU	39	11	68	42.9	4.6	229.1	9.1	1.0	29.3	1.9	0.5	4.1
Greece	GRC	25	19	31	14.9	9.0	20.8	3.9	3.6	4.3	1.1	1.0	1.2
Hungary	HUN	61	61	61	113.2	113.2	113.2	28.1	28.1	28.1	3.5	3.5	3.5
Italy	ITA	53	28	85	32.6	6.9	96.2	15.8	5.3	57.2	2.3	1.5	3.7
Netherlands	NLD	41	31	59	8.2	4.1	14.1	4.7	2.5	8.1	1.4	1.1	1.7
Norway	NOR	47	47	47	16.1	16.1	16.1	5.0	5.0	5.0	1.9	1.9	1.9
Poland	POL	55	12	79	19.4	7.0	52.1	14.1	1.4	29.0	2.2	0.5	2.9
Slovenia	SVN	52	52	52	13.5	13.5	13.5	8.2	8.2	8.2	2.0	2.0	2.0
Spain	ESP	58	37	93	29.3	6.9	93.1	19.7	3.5	102.1	2.4	1.7	3.4
Sweden	SWE	48	40	55	7.7	4.8	10.5	6.0	2.7	8.4	1.6	1.4	1.8
Switzerland	CHE	25	22	31	40.4	36.8	43.8	6.8	3.6	11.3	1.4	1.2	1.7
United States	USA	34	7	81	25.7	1.7	178.7	8.0	1.2	41.4	1.4	0.4	3.5

Source: European data for 2006 from the EC (Urban Audit); US data for 2012 from the OECD (Regional Statistics) and US Census (Population Estimates Program).

For the United States, we used data from the Census Bureau's Population Estimates Program available at <http://www.census.gov/popest/data/cities/totals/2013/>. We used 2012 population data for local units that are incorporated—for example, cities and towns. In three special cases, we used the “consolidated city-county,” where the city and county administration are merged (Indianapolis-Marion county, Indiana; Louisville-Jefferson county, Kentucky; and Nashville-Davidson county, Tennessee). We used the Geographic Correspondence Engine from the Missouri Census Data Center at <http://mcdc.missouri.edu/websas/geocorr12.html> to link towns and cities to the urban areas in the OECD data using a list of counties for each urban area provided by the OECD at <http://www.oecd.org/gov/regional-policy/List-municipalities.xls>.

ACKNOWLEDGMENTS

We thank Lewis Dijkstra (EC) and Monica Brezzi (OECD) for providing additional details about the OECD/EC data, Hugo Poelman (EC) for the preparation of illustrative maps, and Jörg Kalbfuss for excellent help obtaining and managing the various datasets. Comments by the editors—Will Strange, Gilles Duranton, and Vernon Henderson—as well as by Andrew Haughwaut have been particularly helpful. Financial support from the Swiss National Science Foundation (Sinergia grants CRSII1 130648 and 147668) is gratefully acknowledged.

REFERENCES

- Agrawal, D.R., Hoyt, W.H., 2013. State tax differentials, cross-border commuting, and commuting times in multi-state metropolitan areas. University of Georgia and University of Kentucky, Mimeo.
- Ahmad, E., Brosio, G. (Eds.), 2006. Handbook of Fiscal Federalism. Edward Elgar, Cheltenham, UK.
- Alesina, A., Baqir, R., Hoxby, C., 2004. Political jurisdictions in heterogeneous communities. *J. Polit. Econ.* 112 (2), 348–396.
- Allers, M.A., Elhorst, J.P., 2005. Tax mimicking and yardstick competition among local governments in the Netherlands. *Int. Tax Public Financ.* 12, 493–513.
- Andersson, F., Forslid, R., 2003. Tax competition and economic geography. *J. Public Econ. Theory* 5 (2), 279–303.
- Arzaghi, M., Henderson, J.V., 2005. Why countries are fiscally decentralizing. *J. Public Econ.* 89 (7), 1157–1189.
- Baldwin, R.E., Krugman, P., 2004. Agglomeration, integration and tax harmonisation. *Eur. Econ. Rev.* 48 (1), 1–23.
- Baldwin, R.E., Forslid, R., Martin, P., Ottaviano, G.I.P., Robert-Nicoud, F., 2003. *Economic Geography and Public Policy*. Princeton University Press, Princeton, NJ.
- Bardhan, P., Mookherjee, D., 2000. Capture and governance at local and national levels. *Am. Econ. Rev.* 90 (2), 135–139.
- Bayer, P., McMillan, R., Rueben, K., 2004. An equilibrium model of sorting on an urban housing market, NBER Working paper 10865.
- Bayer, P., Ferreira, F., McMillan, R., 2007. A unified framework for measuring preferences for schools and neighborhoods. *J. Polit. Econ.* 115, 588–638.
- Bayer, P.J., McMillan, R., Murphy, A., Timmins, C., 2011. A dynamic model of demand for houses and neighborhoods, NBER Working paper 17250.
- Berry, S., Levinsohn, J., Pakes, A., 1995. Automobile prices in market equilibrium. *Econometrica* 63 (4), 841–890.

- Besley, T., Coate, S., 2003. Centralized versus decentralized provision of local public goods: a political economy approach. *J. Public Econ.* 87, 2611–2637.
- Black, S.E., 1999. Do better schools matter? Parental valuation of elementary education. *Q. J. Econ.* 114, 77–99.
- Blöchliger, H., Rabesona, J., 2009. The fiscal autonomy of sub-central governments: an update, OECD Working papers on Fiscal Federalism 9.
- Boadway, R., 2004. The theory and practice of equalization. *CESifo Econ. Stud.* 50 (1), 211–254.
- Borck, R., Pflüger, M., 2006. Agglomeration and tax competition. *Eur. Econ. Rev.* 50 (3), 647–668.
- Borck, R., Koh, H.-J., Pflüger, M., 2012. Inefficient lock-in and subsidy competition. *Int. Econ. Rev.* 53 (4), 1179–1204.
- Bordignon, M., Cerniglia, F., Revelli, F., 2003. In search of yardstick competition: a spatial analysis of Italian municipality property tax setting. *J. Urban Econ.* 54 (2), 199–217.
- Braid, R., 1996. Symmetrical tax competition with multiple jurisdictions in each metropolitan area. *Am. Econ. Rev.* 86 (5), 1279–1290.
- Braid, R., 2005. Tax competition, tax exporting, and higher government choice of tax instruments for local governments. *J. Public Econ.* 89, 1789–1821.
- Breuilé, M.-L., Zanaj, S., 2013. Mergers in fiscal federalism. *J. Public Econ.* 105, 11–22.
- Brezzi, M., Piacentini, M., Rosina, K., Sanchez-Serra, D., 2012. Redefining urban areas in OECD countries. OECD, Redefining Urban: A New Way to Measure Metropolitan Areas. OECD Publishing, Paris. Available from <http://dx.doi.org/10.1787/9789264174108-en>.
- Brueckner, J.K., 2003. Strategic interaction among governments: an overview of empirical studies. *Int. Reg. Sci. Rev.* 26 (2), 175–188.
- Brueckner, J.K., 2004. Fiscal decentralization with distortionary taxation: Tiebout vs. tax competition. *Int. Tax Public Financ* 11, 133–153.
- Brueckner, J.K., Saavedra, L.A., 2001. Do local governments engage in strategic property-tax competition. *Natl. Tax J.* 56 (2), 203–229.
- Brülhart, M., Jametti, M., 2006. Vertical versus horizontal tax externalities: an empirical test. *J. Public Econ.* 90 (10–11), 2027–2062.
- Brülhart, M., Simpson, H., 2015. Agglomeration economies, taxable rents, and government capture: evidence from a place-based policy. University of Lausanne and University of Bristol, Mimeo.
- Brülhart, M., Schmidheiny, K., 2015. Estimating the rivalness of state-level inward FDI. *J. Reg. Sci.* 55 (1), 139–148.
- Brülhart, M., Trionfetti, F., 2004. Public expenditure, international specialization and agglomeration. *Eur. Econ. Rev.* 48 (4), 851–881.
- Brülhart, M., Jametti, M., Schmidheiny, K., 2012. Do agglomeration economies reduce the sensitivity of firm location to tax differentials? *Econ. J.* 122 (563), 1069–1093.
- Bucovetsky, S., 1991. Asymmetric tax competition. *J. Urban Econ.* 30 (2), 167–181.
- Bucovetsky, S., 2009. An index of tax competition. *Int. Tax Public Financ.* 16 (6), 727–752.
- Bucovetsky, S., Smart, M., 2006. The efficiency consequences of local revenue equalization: tax competition and tax distortions. *J. Public Econ. Theory* 8 (1), 119–144.
- Buettner, T., 2001. Local business taxation and competition for capital: the choice of the tax rate. *Reg. Sci. Urban Econ.* 31 (2), 215–245.
- Buettner, T., 2003. Tax base effects and fiscal externalities of local capital taxation: evidence from a panel of German jurisdictions. *J. Urban Econ.* 54 (1), 110–128.
- Buettner, T., 2006. The incentive effect of fiscal equalization transfers on tax policy. *J. Public Econ.* 90, 477–497.
- Burbidge, J., Cuff, K., 2005. Capital tax competition and returns to scale. *Reg. Sci. Urban Econ.* 35, 353–373.
- Calabrese, S., Epplé, D., Romer, T., Sieg, H., 2006. Local public good provision: voting, peer effects, and mobility. *J. Public Econ.* 90, 959–981.
- Calabrese, S., Epplé, D., Romano, R., 2007. On the political economy of zoning. *J. Public Econ.* 91, 25–49.
- Calabrese, S., Epplé, D., Romano, R., 2012. Inefficiencies from metropolitan political and fiscal decentralization: failures of Tiebout competition. *Rev. Econ. Stud.* 79 (3), 1081–1111.
- Charlot, S., Paty, S., 2007. Market access effect and local tax setting: evidence from French panel data. *J. Econ. Geogr.* 7 (3), 247–263.

- de Bartolome, C.A.M., Ross, S.L., 2003. Equilibria with local governments and commuting: income sorting vs income mixing. *J. Urban Econ.* 54 (1), 1–20.
- de Bartolome, C.A.M., Ross, S.L., 2004. Who's in charge of the central city? The conflict between efficiency and equity in the design of a metropolitan area. *J. Urban Econ.* 56 (3), 458–483.
- de Bartolome, C.A.M., Ross, S.L., 2007. Community income distributions in a metropolitan area. *J. Urban Econ.* 61 (3), 496–518.
- de Mooij, R.A., Ederveen, S., 2003. Taxation and foreign direct investment: a synthesis of empirical research. *Int. Tax Public Financ.* 10 (6), 673–693.
- Devereux, M.P., Lockwood, B., Redoano, M., 2007. Horizontal and vertical indirect tax competition: theory and some evidence from the USA. *J. Public Econ.* 91 (3–4), 451–479.
- Devereux, M.P., Griffith, R., Simpson, H., 2007. Firm location decisions, regional grants and agglomeration externalities. *J. Public Econ.* 91 (3–4), 413–435.
- Dijkstra, L., Poelman, H., 2012. Cities in Europe: the new OECD/EC definition, *Regional Focus*, 1/2012, EU Commission.
- Duranton, G., 2013. Delineating metropolitan areas: measuring spatial labor market networks through commuting patterns. University of Pennsylvania, Mimeo.
- Duranton, G., Gobillon, L., Overman, H.G., 2011. Assessing the effects of local taxation using microgeographic data. *Econ. J.* 121, 1017–1046.
- Edmark, K., Ågren, H., 2008. Identifying strategic interactions in Swedish local income tax policies. *J. Urban Econ.* 63, 849–857.
- Egger, P., Köthenbürger, M., Smart, M., 2010. Do fiscal transfers alleviate business tax competition? Evidence from Germany. *J. Public Econ.* 3–4, 235–246.
- Ellickson, B., 1971. Jurisdictional fragmentation and residential choice. *Am. Econ. Rev.* 61 (2), 334–339.
- Ellison, G., Glaeser, E.L., 1997. Geographic concentration in U.S. manufacturing industries: a Dartboard approach. *J. Polit. Econ.* 105 (5), 889–927.
- Epplé, D., Ferreyra, M.M., 2008. School finance reform: assessing general equilibrium effects. *J. Public Econ.* 92, 1326–1351.
- Epplé, D., Nechyba, T., 2004. Fiscal decentralization. In: Henderson, J.V., Thisse, J.-F. (Eds.), *Handbook of Regional and Urban Economics*, vol. 4. Elsevier, Amsterdam, pp. 2423–2480.
- Epplé, D., Platt, G.J., 1998. Equilibrium and local redistribution in an urban economy when households differ in both preferences and income. *J. Urban Econ.* 43, 23–51.
- Epplé, D., Sieg, H., 1999. Estimating equilibrium models of local jurisdictions. *J. Polit. Econ.* 107 (4), 645–681.
- Epplé, D., Romer, T., Sieg, H., 2001. Interjurisdictional sorting and majority rule: an empirical analysis. *Econometrica* 69 (6), 1437–1466.
- Epplé, D., Gordon, B., Sieg, H., 2010. Drs. Muth and Mills meet Dr. Tiebout: integrating location-specific amenities into multi-community equilibrium models. *J. Reg. Sci.* 50 (1), 381–400.
- Epplé, D., Peress, M., Sieg, H., 2010. Identification and semiparametric estimation of equilibrium models of local jurisdictions. *Am. Econ. J. Microecon.* 2 (4), 195–220.
- Epplé, D., Romano, R., Sieg, H., 2012. The intergenerational conflict over the provision of public education. *J. Public Econ.* 96 (3), 255–268.
- Eugster, B., Parchet, R., 2014. Culture and taxes: towards identifying tax competition. University of St. Gallen and University of Lugano, Mimeo.
- Feld, L.P., Kirchgässner, G., 2001. Income tax competition at the state and local level in Switzerland. *Reg. Sci. Urban Econ.* 31, 181–213.
- Fernandez, G.E., 2005. A note on tax competition in the presence of agglomeration economies. *Reg. Sci. Urban Econ.* 35, 837–847.
- Frey, B.S., Eichenberger, R., 1996. To harmonize or to compete? That's not the question. *J. Public Econ.* 60 (3), 335–349.
- Friedman, J., 1981. A conditional logit model of the role of local public services in residential choice. *Urban Stud.* 18, 347–358.
- Fuest, C., Huber, B., Mintz, J., 2005. Capital mobility and tax competition. *Found. Trends Microecon.* 1 (1), 1–62.
- Gabszewicz, J., Tarola, O., Zanaj, S., 2013. Migration, wages and fiscal competition. University of Luxembourg, Technical Report 13–19, CREA.

- Gaigné, C., Riou, S., 2007. Globalization, asymmetric tax competition, and fiscal equalization. *J. Public Econ. Theory* 9 (5), 901–925.
- Gaigné, C., Riou, S., Thisse, J.-F., 2013. How to make the metropolitan area work? Neither big government, nor laissez-faire, CORE Discussion Paper No. 2013/65.
- Glaeser, E.L., 2013. Urban public finance. In: Auerbach, A.J., Chetty, R., Feldstein, M., Saez, E. (Eds.), *Handbook of Public Economics*, vol. 5. Elsevier, Amsterdam, pp. 195–256.
- Goodspeed, T.J., 1989. A re-examination of the use of ability to pay taxes by local governments. *J. Public Econ.* 38, 319–342.
- Gordon, R.H., Hines, J., 2002. International taxation. In: Auerbach, A.J., Martin, F. (Eds.), *Handbook of Public Economics*, vol. 4. North-Holland, Amsterdam, pp. 1935–1995.
- Gordon, N., Knight, B., 2009. A spatial merger estimator with an application to school district consolidation. *J. Public Econ.* 93 (5–6), 752–765.
- Gravel, N., Oddou, R., 2014. The segregative properties of endogenous jurisdiction formation with a land market. *J. Public Econ.* 117, 15–27.
- Hansen, N.A., Kessler, A.S., 2001. The political geography of tax h(e)avens and tax hells. *Am. Econ. Rev.* 91, 1103–1115.
- Haughwout, A.F., Inman, R.P., 2009. How should suburbs help their central cities? Growth- and welfare-enhancing intrametropolitan fiscal distributions. *Ann. Am. Acad. Polit. Soc. Sci.* 626, 39–52.
- Haughwout, A.F., Inman, R.P., Henderson, J.V., 2002. Should suburbs help their central city? Brookings-Wharton Pap. Urban Aff. 2002, 45–94.
- Haughwout, A.F., Inman, R.P., Craig, S., Luce, T., 2004. Local revenue hills: evidence from four U.S. cities. *Rev. Econ. Stat* 86 (2), 570–585.
- Hauptmeier, S., Mittermaier, F., Rincke, J., 2012. Fiscal competition over taxes and public inputs: theory and evidence. *Reg. Sci. Urban Econ.* 42 (3), 407–419.
- Helsley, R.W., 2004. Urban political economics. In: Henderson, J.V., Thisse, J.-F. (Eds.), *Handbook of Regional and Urban Economics*, vol. 4. North-Holland, Amsterdam, pp. 2381–2421.
- Henderson, J.V., 1985. The Tiebout model: bring back the entrepreneurs. *J. Polit. Econ.* 93 (2), 248–264.
- Henderson, J.V., Storeygard, A., Weil, D.N., 2012. Measuring economic growth from outer space. *Am. Econ. Rev.* 102, 994–1028.
- Hines, J.R., 1999. Lessons from behavioral responses to international taxation. *Natl. Tax J.* 52 (2), 305–322.
- Hochman, O., Pines, D., Thisse, J.-F., 1995. On the optimal structure of local governments. *Am. Econ. Rev.* 85 (5), 1224–1240.
- Hodler, R., Schmidheiny, K., 2006. How fiscal decentralization flattens progressive taxes. *FinanzArchiv Public Financ. Anal.* 62 (2), 281–304.
- Hoxby, C.M., 2001. All school finance equalizations are not created equal. *Q. J. Econ.* 116 (4), 1189–1231.
- Hoyt, W.H., 2001. Tax policy coordination, vertical externalities, and optimal taxation in a system of hierarchical governments. *J. Urban Econ.* 50, 491–561.
- Hühnerbein, O., Seidel, T., 2010. Intra-regional tax competition and economic geography. *World Econ.* 33, 1042–1051.
- International Monetary Fund, 2001. *Government Finance Statistics Manual 2001 (GFSM 2001)*. Available online at <https://www.imf.org/external/pubs/ft/gfs/manual/>.
- Ioannides, Y., 2013. *From Neighborhoods to Nations*. Princeton University Press, Princeton.
- Janeba, E., Osterloh, S., 2013. Tax and the city—a theory of local tax competition. *J. Public Econ.* 106, 89–100.
- Jofre-Monseny, J., 2013. Is agglomeration taxable? *J. Econ. Geogr.* 13 (1), 177–201.
- Jofre-Monseny, J., Solé-Ollé, A., 2012. Which communities should be afraid of mobility? The effects of agglomeration economies on the sensitivity of employment location to local taxes. *Reg. Sci. Urban Econ.* 42, 257–268.
- Kächelein, H., 2014. Asymmetric tax competition and choice of tax rate: commuting as an explanation for tax differentials. *Int. Tax Public Financ.* 21 (1), 50–65.
- Kanbur, R., Keen, M.J., 1993. Jeux Sans Frontières: tax competition and tax coordination when countries differ in size. *Am. Econ. Rev.* 83, 877–892.
- Keen, M.J., Konrad, K., 2013. The theory of international tax competition and coordination. In: Auerbach, A.J., Raj, C., Martin, F., Emmanuel, S. (Eds.), *Handbook of Public Economics*, vol. 5. Elsevier, Amsterdam, pp. 257–328.

- Keen, M.J., Kotsogiannis, C., 2002. Does federalism lead to excessively high taxes? *Am. Econ. Rev.* 92, 363–370.
- Keen, M.J., Kotsogiannis, C., 2004. Tax competition in federations and the welfare consequences of decentralization. *J. Urban Econ.* 56 (3), 397–407.
- Kessler, A.S., 2014. Communication in federal politics: universalism, policy uniformity, and the optimal allocation of fiscal authority. *J. Polit. Econ.* 122 (4), 766–805.
- Kind, H.J., Knarvik, K.H.M., Schjelderup, G., 2000. Competing for capital in a ‘lumpy’ world. *J. Public Econ.* 78 (3), 253–274.
- Kirchgässner, G., Pommerehne, W., 1996. Tax harmonization and tax competition in the European union: lessons from Switzerland. *J. Public Econ.* 60, 351–371.
- Kleven, H.J., Landais, C., Saez, E., 2013. Taxation and international migration of superstars: evidence from the European football market. *Am. Econ. Rev.* 103 (5), 1892–1924.
- Kleven, H.J., Landais, C., Saez, E., Schultz, E., 2014. Migration and wage effects of taxing top earners: evidence from the foreigners’ tax scheme in Denmark. *Q. J. Econ.* 129 (1), 333–378.
- Koh, H.-J., Riedel, N., Böhm, T., 2013. Do governments tax agglomeration rents? *J. Urban Econ.* 75, 92–106.
- Konrad, K.A., Kovenock, D., 2009. Competition for FDI with vintage investment and agglomeration advantages. *J. Int. Econ.* 79 (2), 230–237.
- Köthenbürger, M., 2002. Tax competition and fiscal equalization. *Int. Tax Public Financ.* 9, 391–408.
- Köthenbürger, M., Lockwood, B., 2010. Does tax competition really promote growth? *J. Econ. Dyn. Control* 34 (2), 191–206.
- Krostrup, S., 2008. Standard tax competition and increasing returns. *J. Public Econ. Theory* 10 (4), 547–561.
- Kuminoff, N.V., Smith, V.K., Timmins, C., 2013. The new economics of equilibrium sorting and its transformational role for policy evaluation. *J. Econ. Lit.* 51 (4), 1007–1062.
- Liebig, T., Puhani, P.A., Sousa-Poza, A., 2007. Taxation and internal migration—evidence from the Swiss census using community-level variation in income tax rates. *J. Reg. Sci.* 47 (4), 807–836.
- Ludema, R.D., Wooton, I., 2000. Economic geography and the fiscal effects of regional integration. *J. Int. Econ.* 52 (2), 331–357.
- Luthi, E., Schmidheiny, K., 2014. The effect of agglomeration size on local taxes. *J. Econ. Geogr.* 14 (2), 265–287.
- Marceau, N., Mongrain, S., Wilson, J.D., 2010. Why do most countries set high tax rates on capital? *J. Int. Econ.* 80 (2), 249–259.
- McFadden, D., 1974. Conditional logit analysis of qualitative choice behavior. In: Zarembka, P. (Ed.), *Frontiers in Econometrics*. Academic Press, New York, pp. 105–142.
- McFadden, D., 1978. Modelling the choice of residential location. In: Karlqvist, A. et al., (Eds.), *Spatial Interaction Theory and Planning Models*. North-Holland, Amsterdam, pp. 75–96.
- Meltzer, A., Richard, S., 1981. A rational theory of the size of government. *J. Polit. Econ.* 89, 914–927.
- Nechyba, T.J., Strauss, R.P., 1998. Community choice and local public services: a discrete choice approach. *Reg. Sci. Urban Econ.* 28, 51–73.
- Oates, W.E., 1972. *Fiscal Federalism*. Harcourt Brace, New York, pp. 75–96.
- OECD, 1999. *Taxing Powers of State and Local Government*, OECD Tax Policy Studies No. 1, Paris.
- OECD, 2013. *OECD Regions at a Glance 2013*. OECD Publishing, Paris.
- OECD, 2013b. Definition of Functional Urban Areas (FUA) for the OECD Metropolitan Database. Available online at, <http://www.oecd.org/gov/regional-policy/Definition-of-Functional-Urban-Areas-for-the-OECD-metropolitan-database.pdf>.
- Parchet, R., 2014. Are local tax rates strategic complements or strategic substitutes? University of Lugano, Mimeo.
- Peralta, S., van Ypersele, T., 2005. Factor endowments and welfare levels in an asymmetric tax competition game. *J. Urban Econ.* 57 (2), 258–274.
- Pieretti, P., Zanaj, S., 2011. On tax competition, public goods provision, and jurisdictions’ size. *J. Int. Econ.* 84, 124–130.
- Redoano, M., 2010. Does centralization affect the number and size of lobbies? *J. Public Econ. Theory* 12 (3), 407–435.
- Riou, S., 2006. Transfer and tax competition in a system of hierarchical governments. *Reg. Sci. Urban Econ.* 36, 249–269.

- Roberts, K.W.S., 1977. Voting over income tax schedules. *J. Public Econ.* 8, 329–340.
- Romer, T., Rosenthal, S.S., Ross, A., 2014. Tax avoidance and business location in a state border model. *J. Urban Econ.* *forthcoming*.
- Romer, T., 1975. Individual welfare, majority voting, and the properties of a linear income tax. *J. Public Econ.* 7, 163–168.
- Rosenthal, S.S., Strange, W.C., 2004. Evidence on the Nature and Sources of Agglomeration Economies. In: Henderson, J.V., Thisse, J.F. (Eds.), *Handbook of Regional and Urban Economics*, vol. 4. Elsevier, Amsterdam, pp. 2119–2171.
- Ross, S., Yinger, J., 1999. Sorting and voting: a review of the literature on urban public finance. In: Cheshire, P., Mills, E.S. (Eds.), *Handbook of Regional and Urban Economics*, vol. 3. North-Holland, Amsterdam, pp. 2001–2060.
- Rozenfeld, H.D., Rybski, D., Andrade, J.S., Batty, M., Stanley, H.E., Makse, H.A., 2008. Laws of population growth. *Proc. Natl. Acad. Sci. U.S.A.* 105, 18702–18707.
- Rozenfeld, H.D., Rybski, D., Gabaix, X., Makse, H.A., 2011. The area and population of cities: new insights from a different perspective on cities. *Am. Econ. Rev.* 101, 2205–2225.
- Schmidheiny, K., 2006. Income segregation and local progressive taxation: empirical evidence from Switzerland. *J. Public Econ.* 90 (3), 429–458.
- Schmidheiny, K., 2006. Income segregation from local income taxation when households differ in both preferences and incomes. *Reg. Sci. Urban Econ.* 36 (2), 270–299.
- Schmidheiny, K., Brühlhart, M., 2011. On the equivalence of location choice models: conditional logit, nested logit and poisson. *J. Urban Econ.* 69 (2), 214–222.
- Solé-Ollé, A., 2006. Expenditure spillovers and fiscal interactions: empirical evidence from local governments in Spain. *J. Urban Econ.* 59 (1), 32–53.
- Stegarescu, D., 2006. Decentralised Government in an Integrating World. *Physica*, Heidelberg.
- Tiebout, C., 1956. A pure theory of local expenditures. *J. Polit. Econ.* 64, 416–424.
- Westhoff, F., 1977. Existence of equilibria in economies with a local public good. *J. Econ. Theory* 14, 84–112.
- Wildasin, D.E., 1988. Nash equilibrium in models of tax competition. *J. Public Econ.* 35, 229–240.
- Wilson, J.D., 1986. A theory of interregional tax competition. *J. Urban Econ.* 19, 296–315.
- Wilson, J.D., 1991. Tax competition with interregional differences in factor endowments. *Reg. Sci. Urban Econ.* 21, 423–451.
- Wilson, J.D., 1999. Theories of tax competition. *Natl. Tax J.* 52 (2), 269–304.
- Wrede, M., 1997. Tax competition and federalism: the underprovision of local public goods. *Finanzarchiv* 54, 494–515.
- Wrede, M., 2014. Agglomeration, tax competition, and fiscal equalization. *Int. Tax Public Financ.* *forthcoming*.
- Young, C., Varner, C., 2011. Millionaire migration and state taxation of top incomes: evidence from a natural experiment. *Natl. Tax J.* 64 (2), 255–283.
- Zipf, G.K., 1949. *Human Behavior and the Principle of Least Effort*. Addison-Wesley, Cambridge.
- Zodrow, G.R., Mieszkowski, P., 1986. Pigou, Tiebout, property taxation and the underprovision of local public goods. *J. Urban Econ.* 19, 356–370.

CHAPTER 18

Place-Based Policies

David Neumark*, Helen Simpson†

*UCI, NBER, and IZA, Irvine, CA, USA

†University of Bristol, CMPO, OUCBT and CEPR, Bristol, UK

Contents

18.1. Introduction	1198
18.2. Theoretical Basis for Place-Based Policies	1206
18.2.1 Agglomeration economies	1206
18.2.2 Knowledge spillovers and the knowledge economy	1208
18.2.3 Industry localization	1209
18.2.4 Spatial mismatch	1210
18.2.5 Network effects	1211
18.2.6 Equity motivations for place-based policies	1212
18.2.7 Summary and implications for empirical analysis	1213
18.3. Evidence on Theoretical Motivations and Behavioral Hypotheses Underlying Place-Based Policies	1215
18.3.1 Evidence on agglomeration economies	1215
18.3.2 Is there spatial mismatch?	1217
18.3.3 Are there important network effects in urban labor markets?	1219
18.4. Identifying the Effects of Place-Based Policies	1221
18.4.1 Measuring local areas where policies are implemented and economic outcomes in those areas	1222
18.4.2 Accounting for selective geographic targeting of policies	1222
18.4.3 Identifying the effects of specific policies when areas are subject to multiple interventions	1225
18.4.4 Accounting for displacement effects	1225
18.4.5 Studying the effects of discretionary policies targeting specific firms	1226
18.4.6 Relative versus absolute effects	1229
18.5. Evidence on Impacts of Policy Interventions	1230
18.5.1 Enterprise zones	1230
18.5.1.1 <i>The California enterprise zone program</i>	1230
18.5.1.2 <i>Other recent evidence for US state-level and federal programs</i>	1237
18.5.1.3 <i>Evidence from other countries</i>	1246
18.5.1.4 <i>Summary of evidence on enterprise zones</i>	1249
18.5.2 Place-based policies that account for network effects	1250
18.5.3 Discretionary grant-based policies	1252
18.5.3.1 <i>Summary of evidence on discretionary grants</i>	1259
18.5.4 Clusters and universities	1261
18.5.4.1 <i>Clusters policies</i>	1261
18.5.4.2 <i>Universities</i>	1264
18.5.4.3 <i>Summary of evidence on clusters and universities</i>	1267

18.5.5 Infrastructure investment and other regional policies	1268
18.5.6 Community development and locally led initiatives	1275
18.6. Unanswered Questions and Research Challenges	1279
Acknowledgments	1282
References	1282

Abstract

Place-based policies commonly target underperforming areas, such as deteriorating downtown business districts and disadvantaged regions. Principal examples include enterprise zones, European Union Structural Funds, and industrial cluster policies. Place-based policies are rationalized by various hypotheses in urban and labor economics, such as agglomeration economies and spatial mismatch—hypotheses that entail market failures and often predict overlap between poor economic performance and disadvantaged residents. The evidence on enterprise zones is very mixed. We need to know more about what features of enterprise zone policies make them more effective or less effective, who gains and who loses from these policies, and how we can reconcile the existing findings. Some evidence points to positive benefits of infrastructure expenditure and also investment in higher education and university research—likely because of the public-goods nature of these policies. However, to better guide policy, we need to know more about what policies create self-sustaining longer run gains.

Keywords

Place-based policies, Employment, Enterprise zones, Discretionary grants, Higher education, Industrial clusters, Infrastructure

JEL Classification Codes

R12, R38, J68, H25

18.1. INTRODUCTION

Broadly speaking, place-based policies refer to government efforts to enhance the economic performance of an area within its jurisdiction, typically in the form of more job opportunities and higher wages. Best known, perhaps, are place-based policies that target underperforming areas, such as deteriorating downtown business districts or, within the European Union, relatively disadvantaged areas eligible for regional development aid. Alternatively, place-based policies may seek to enhance even further the economic performance of areas that are already doing well.

Ladd (1994) distinguished a subset of place-based policies or strategies that she labeled “place-based people strategies.” These are policies that are geographically targeted, but with the intent and structure of helping disadvantaged residents in them—for example, enterprise zone programs that seek to create jobs in or near areas where poor people live and job prospects are weak. In contrast, some place-based policies target areas irrespective of whether there are disadvantaged people living in those areas, or even many people at

all, such as efforts to revitalize a downtown business district including real-estate development or initiatives to help strengthen an industrial cluster in a region.

Place-based people strategies, in particular, can be contrasted with “people-based” policies that try to help the disadvantaged without regard to where they live or how concentrated they are. Examples include welfare and working tax credits (such as the earned income tax credit in the United States). People-based policies are the more traditional purview of public finance and are not covered in this chapter. Rather, the chapter focuses on a wide range of place-based policies—including pure place-based policies and place-based people policies.

Place-based policies that also focus on people can be categorized as direct or indirect. Direct forms of place-based policies seek to increase economic activity and strengthen labor markets where disadvantaged people currently live, while indirect policies may instead seek to increase access of those people to locations where labor markets are stronger. Enterprise zones can be viewed as direct, since they typically create incentives for hiring, or economic activity more generally, in or near areas where disadvantaged people live. The Gautreaux Project and Moving to Opportunity program in the United States, as well as transportation-based policies intended to increase access to jobs outside of areas where the disadvantaged tend to reside (in the United States, the urban core)—that is, intended to reduce spatial mismatch—are examples of indirect policies. However, this chapter focuses on direct policies.¹

Place-based policies targeting the disadvantaged, including indirect policies, are often rationalized in part by hypotheses that seek to explain the overlap between areas with poor economic performance and disadvantaged residents, coupled with market failures of one form or another. The standard arguments considered in the urban economics literature to rationalize pure place-based policies are generally efficiency arguments pertaining to the existence of agglomeration externalities. But this literature also calls into question whether policies that aim to stimulate economic activity in one place rather than another deliver any aggregate benefits and whether place-based people policies will ultimately help those individuals they target.

In our view, other market imperfections that have been highlighted in the labor economics literature may also justify place-based policies of both types. One is the spatial mismatch hypothesis, wherein minorities or low-skilled workers in some urban areas may face long-term disadvantage spurred by declines in employment opportunities as manufacturing jobs left the cities, coupled with housing discrimination or other constraints that restrict their mobility to locations with better employment opportunities. A second is positive externalities stemming from network effects, whereby employment

¹ There are many excellent summaries of the details of both the Gautreaux Project and MTO program designs, and there are a number of comprehensive reviews of findings of studies of either or both programs; see, e.g., [Duncan and Zuberi \(2006\)](#), [Rosenbaum and Zuberi \(2010\)](#), and [Ludwig et al. \(2013\)](#).

of residents can help other residents find jobs (e.g., [Hellerstein et al., 2011](#)). Either the externalities from network effects or the mobility constraints implied by spatial mismatch can potentially justify geographically targeted policies to increase employment. This chapter reviews evidence on these labor-market hypotheses that can potentially rationalize place-based policies, with a more cursory discussion of the standard urban economics hypotheses regarding agglomeration and spillovers, on which plenty of work already exists.

The majority of the chapter focuses on the research evidence on impacts of place-based policies and discusses issues arising in the empirical identification of causal effects in this setting.² In the remainder of this section, we provide more details on the types of place-based policies we consider and emphasize the intended recipients and the stated goals of these policies. Later in the chapter, in both the context of the theoretical basis for these interventions and the evidence on their effects, we consider whether these goals are met. Due to space constraints, we limit our coverage throughout to place-based policies in the United States and in Europe. This focus allows us to contrast evidence on similar types of policies implemented in both locations and, where the evaluation literature has examined comparable outcomes using similar empirical approaches, enables us to draw conclusions that are more general. In turn, this means that we necessarily exclude interventions in developing countries, such as Special Economic Zones in China (see [Alder et al., 2013](#), and [Wang, 2013](#)) and India's National Investment and Manufacturing Zones.

We also exclude policies that result from political or fiscal decentralization and that apply across whole jurisdictions (and, therefore, without regard to the characteristics of the areas where the incentives apply or the people who live in them), rather than to areas *within* a jurisdiction. Examples include discretionary programs, such as the Michigan Economic Growth Authority (MEGA), which provides tax credits to businesses in the state's export industries ([Bartik and Erickcek, 2010](#)), and broader policies on which jurisdictions may compete to attract businesses. There is a large literature on tax competition between areas to attract firms—such as through research and development tax credits (e.g., [Wilson, 2009](#); [Chang, 2013](#)), covered by [Brühlhart et al. \(2015\)](#). And states and cities are often viewed as competing on a number of dimensions including taxes, regulations, and quality of life, which are often captured and summarized in business climate indexes ([Kolko et al., 2013](#)). [Bartik \(2003\)](#) also discussed the potential role of customized economic development services for businesses. Because these kinds of policies and dimensions of competition fall outside of the usual definition of place-based policies that

² [Kline and Moretti \(2014a\)](#) provided a very useful complementary review article on place-based policies that focuses largely on a theoretical discussion of the welfare economics of local economic development programs, with a very limited discussion of the evidence. In contrast, our goal is to provide a comprehensive overview and evaluation of the evidence base.

try to reallocate economic activity across areas within a jurisdiction or stimulate activity in very specific areas within a jurisdiction, they are not covered in this chapter.

The place-based policy that has attracted the most attention from researchers is enterprise zones. In the United States, these exist at both the federal and state levels.³ For example, under the federal Empowerment Zone Program in the United States, authorized in 1993, local governments could submit proposals for zones made up of relatively poor, high-unemployment Census tracts.⁴ The federal Enterprise Community program, also authorized in 1993, had the same criteria. Far more Enterprise Communities than Empowerment Zones were created. The former had much smaller benefits—grants of just under \$3 million versus \$100 million (\$40 million) for urban (rural) Empowerment Zones (US Government Accountability Office, 2006)—and much less generous hiring credits.⁵ Spending through 2000 in the first round of the federal enterprise zone program totaled nearly \$400 million in block grants and \$200 million in employment credits. Federal expenditures via hiring credits and block grants for the first 6 years of the program are estimated at about \$850 per zone resident.

There is a plethora of state enterprise zones programs in the United States—40 as of 2008 (Ham et al., 2011). These vary in size (some even cover the entire state!), the number of zones in each state, and the benefits available. As an example of targeting, however, consider the case of California, whose state enterprise zone program has been studied most extensively. In California, enterprise zones are supposed to be areas with job creation potential that are near and can overlap with Targeted Employment Areas (TEAs), consisting of Census tracts where more than half the population earns less than 80% of median area income, according to the 1980 Census.⁶ The most significant benefit is a hiring credit to businesses located in zones. A worker living in a TEA qualifies for the hiring credit regardless of their characteristics. Clearly, both federal enterprise zone programs and this state program (and the same is true of many others) target areas based on the characteristics of people who live in them.

³ Bartik (2003) noted that earlier related programs focusing on distressed communities include “Urban Renewal” in the 1940s and 1950s, “Model Cities” during the War on Poverty, and Community Development Block Grants.

⁴ All tracts in the zone had to have poverty rates above 20%, with at least 90% of tracts above 25%, and 50% of tracts above 35%. In addition, unemployment rates in each tract had to exceed the 1990 national average of 6.3% (US Government Accountability Office, 2006; Busso et al., 2013).

⁵ The Enterprise Communities were created among applicant areas that did not receive Empowerment Zone designation, leading Busso et al. (2013) to characterize the Enterprise Communities as “consolation prizes.” The rejected status of these areas figures prominently in research discussed later. In 2000, an additional program (Renewal Communities), with related but different criteria, was established, offering a hiring credit and other benefits. See http://portal.hud.gov/hudportal/documents/huddoc?id=19132_actof2000.pdf (viewed 11 July 2013).

⁶ Other studies describe similar types of criteria for federal programs (e.g., Hanson, 2009) and programs in other states (e.g., Lynch and Zax, 2011).

Enterprise zone policies are also used in some European countries. France introduced an enterprise zone program in 1997 (Zones Franches Urbaines (ZFU)), targeting municipalities or groups of municipalities facing acute unemployment, as well as high poverty and other economic challenges. The criteria used to define these areas included population, population aged under 25, unemployment rate, fraction of the population with no skills, and the fiscal potential of the area, which is related to income (Gobillon et al., 2012). The policy offered relief on property taxes, corporate income taxes, and wages and aimed to increase local employment by making the wage tax relief conditional on hiring at least 20% of employees locally. The United Kingdom ran a program of enterprise zones from 1981 to the mid-1990s covering areas of derelict industrial land in locations that had been hit by industrial decline (Papke, 1993) and that aimed to create local jobs through new businesses and inward foreign direct investment. The policy offered incentives for business investment including more generous tax allowances for capital investment, exemptions from business rates (a local tax on commercial property), and relaxation of planning regulations.⁷

A quite different type of place-based policy that also targets economically disadvantaged areas is a larger scale government effort to help economic development through infrastructure investment. A prime example is the Tennessee Valley Authority, a federal initiative to modernize the economy of the Tennessee Valley Region, encompassing most of Tennessee and parts of Kentucky, Alabama, and Mississippi. The program entailed large public infrastructure spending with an emphasis on hydroelectric dams to generate power sold locally to encourage manufacturing and other spending on, for example, schools and flood controls (Kline and Moretti, 2014b). Another example is the Appalachian Regional Commission (discussed by Glaeser and Gottlieb, 2008) that provided assistance focused on transportation for a large swath of states extending from Mississippi to New York, beginning in 1963.

Within the European Union (EU), Structural Funds—comprising the European Regional Development Fund (ERDF) and European Social Fund (ESF)—support a wide range of initiatives aimed at economic development and increasing labor-market participation and skills; these policies also generally target disadvantaged areas.⁸ Expenditure under the ERDF can include investment in transport or telecommunications infrastructure or investment linked to innovation, the environment, or energy. The ESF is used to provide funding for programs aimed at reducing unemployment, increasing human capital, and increasing social integration of disadvantaged groups. The bulk of Structural Funds expenditure flows to the so-called Objective 1 areas. These are regions within

⁷ During the 1980s, Spain implemented a reindustrialization zone policy and Belgium a program of employment zones. France also operated an earlier enterprise zone policy in the late 1980s.

⁸ For 2007–2013, expenditure on Structural Funds was 278 billion euros, a significant fraction of the European Community budget (see http://europa.eu/legislation_summaries/glossary/structural_cohesion_fund_en.htm, viewed 6 January 2014).

EU member states with GDP per capita less than 75% of the European Community average. For 2007–2013, many new member countries such as Poland and Romania were entirely classified as Objective 1 areas. Other examples include peripheral regions such as in southern Italy, southern Spain, and Portugal and some lagging regions in the United Kingdom and (former East) Germany.

Under EU legislation, European governments can also offer subsidies to private-sector firms within these areas. Since the 1970s, the United Kingdom has run a number of discretionary grant schemes (e.g., Regional Selective Assistance, Regional Development Grants, and Enterprise Grants) that subsidize new capital investment with explicit aims of creating or safeguarding jobs and attracting inward investment. The grants are available in designated, relatively disadvantaged “Assisted Areas” within the United Kingdom, with area eligibility determined by GDP per capita and unemployment rate indicators relative to the EU average. The subsidy rate allowable varies with area characteristics, with Objective 1 areas eligible for the highest subsidy rates.⁹ A similar grant program has been in operation in France (the *Prime d’Aménagement du Territoire*), and the Italian government operates a scheme known as Law 488. Although on paper the direct recipients of these subsidies are businesses, the ultimate intended beneficiaries are individuals residing in these lagging regions; hence, these programs have a people-based flavor.

There are other European place-based policies, directly aimed at firms, which do not necessarily have a people-based component: for example, support for industrial clusters outside of relatively deprived areas. The current UK enterprise zone policy, which began in 2011, aims to increase new business start-ups and growth and to create new jobs. Within England, there are now 24 designated areas not only offering some of the same tax incentives as the previous scheme but also aiming to promote clusters of businesses within the same industrial sector and emphasizing location-specific amenities including access to transport infrastructure such as rail and ports. The motivation for cluster policies often comes from evidence on productivity benefits arising from industry localization or on the observed colocation of some high-tech clusters with higher education institutions. In Sweden, the government has explicitly tried to use the location of new universities as a regional policy tool to both increase local labor force skills and potentially exploit knowledge spillovers from university research as a means of attracting private-sector activity to an area and boosting local productivity.

As this discussion suggests, there is a large variety of policies that can be considered under the general rubric of place-based policies. [Table 18.1](#) provides summary information on the general types of place-based policies that exist, as well as some details on specific examples. Some have been mentioned already, and others will be discussed in the sections that follow.

⁹ Eligible areas are revised every 7 years. The precise set of economic indicators and geographic units used to define eligible areas have varied over time.

Table 18.1 Place-based policies

Type of policy	Enterprise zone	Business development, attraction, and retention	Cluster promotion	Infrastructure investment		Discretionary grants	Community development and locally led initiatives
Specific examples	California enterprise zone program; US Federal Empowerment Zones; US Federal Enterprise Communities; French enterprise zones	UK enterprise zones (2011)	French Local Productive Systems; Bavarian High-Tech Offensive	Tennessee Valley Authority; Appalachian Regional Commission	EU Structural Funds: European Regional Development Fund (ERDF), European Social Fund (ESF)	UK Regional Selective Assistance; Italian Law 488	Low-Income Housing Tax Credit; redevelopment areas; New Markets Tax Credit
Policy goals	Job creation	New business creation; job creation; industry clustering	Increase collaboration and cooperation between firms, and between firms and public-sector research institutions	Economic modernization	ERDF: economic development ESF: increased labor-market participation	Job creation and safeguarding; inward investment	Affordable housing; urban redevelopment; economic development
Targeting	Areas with higher concentrations of poverty, unemployment	New businesses within government-designated areas	France: no restriction on local areas that could participate Bavaria: whole state	Poor areas of region	Areas with relatively low GDP per capita/high unemployment relative to the EU average	Areas with relatively low GDP per capita/high unemployment relative to the EU average	Low-income neighborhoods or low-income housing units
Incentives	Hiring tax credits; corporate and personal income tax credits; sales	Reduced business rates; relaxed planning regulation;	France: subsidy for a project, e.g., to boost exports of	Reduced electricity rates; other infrastructure improvements	ERDF: transport, telecommunications infrastructure, and investment linked to innovation or energy	Subsidy on new investment in physical capital by firms, linked to jobs targets	Tax credits to investors or real-estate developers; tax

	and use tax credits; tax-exempt financing; community block grants; property, corporate, and wage tax relief	enhanced capital allowances in some cases	participating firms Bavaria: access to public research facilities, venture capital funding, and science parks France: businesses in a common industry Bavaria: targeted five high-tech sectors		ESF: training programs		increment financing
Recipients of support	Mainly businesses; sometimes workers; communities	Businesses		Broad	Broad	Primarily manufacturing businesses	Real-estate developers or other businesses

18.2. THEORETICAL BASIS FOR PLACE-BASED POLICIES

In assessing the welfare effects of place-based policies, theory highlights some important factors, which in turn can be used to direct empirical analysis of policy effects. Key questions include the following: Can policy exploit agglomeration externalities or solve other market failures to generate long-term gains for targeted areas? If so, does intervention come at a cost to other areas, and are there any aggregate national benefits of location-specific interventions? Does policy that targets specific places create distortions to capital and labor mobility, lowering efficiency by reducing incentives of firms or individuals to move to other more productive locations? And how does geographic mobility affect outcomes for those originally resident in the targeted areas, as well as the eventual incidence of a place-based policy? In short, can intervention be justified, and what potential effects of place-based policies should empirical analysis aim to identify?

Before considering potential efficiency rationales for intervention, it is worth starting from the benchmark of the absence of market failures. With perfect labor mobility combined with inelastic housing supply in the targeted area, theory implies that, as a result of in-migration and increased demand for housing, landowners benefit from a location-specific policy, rather than local residents, with the benefits being capitalized into rents. With less than perfect labor mobility, local residents may benefit, but these benefits should be weighed against any costs to nontargeted areas and the deadweight costs of taxation. Place-based policies may be justified in the context of market failures that have a spatial dimension. In the rest of this section, we outline possible arguments why place-based policies may help overcome specific market imperfections or take advantage of externalities and consider the case for redistribution or equity-motivated policies that target disadvantaged areas and not just disadvantaged people.

18.2.1 Agglomeration economies

The efficiency-related argument for place-based policies that is most central to urban economics is that there exist agglomeration economies, through which the dense population of urban areas has an independent effect on the productivity of resources. Agglomeration economies may arise via a number of mechanisms, which [Duranton and Puga \(2004\)](#) categorized as “sharing, matching, and learning.” [Moretti \(2010\)](#) emphasized the role of thick labor markets, which can lead to better worker–firm matches, inducing more investment by workers and firms. Thick labor markets can also provide better insurance against local demand shocks by reducing the risk or cost of unemployment, which can act as a compensating differential that lowers labor costs. Moretti also emphasized thick markets for intermediate inputs, especially those that are specialized and nontradable. Examples are professional services, such as computer

programming, legal support, and venture capital.¹⁰ If a firm needs these inputs from other companies, it has an incentive to locate in a city with other firms that use the same inputs.

Agglomeration economies imply positive externalities, because bringing additional people or firms to an urban area increases the productivity of other individuals or firms in that area, but these gains are not captured by those deciding whether to move to that location. Thus, there may be a rationale to subsidize in-migration or growth, to raise the private returns closer to the social returns.¹¹ Moretti (2010) argued that the rationale for place-based policies to exploit agglomeration economies may be quite strong, especially in a dynamic setting with multiple equilibria, in which externalities can generate benefits from drawing economic activity to any single one of a set of ex ante similar locations. In such cases, the gains from moving from a low-employment, low-density equilibrium to a high one at a particular location may far exceed the costs of the policy (such as a temporary, but large-scale intervention as in Kline, 2010). Hence, there may be a case for place-based policy to jump-start growth in a specific area. Nevertheless, two further questions need to be addressed: Which areas should policy target? And will the gains to those areas be offset by losses to others?

As Glaeser and Gottlieb (2008) emphasized, in choosing *between* locations in which to encourage growth, policymakers should do so in areas where the elasticity of productivity with respect to agglomeration is higher (which may well not be the most deprived areas), exploiting spatial variation in the relationship between productivity and size or density. In practice, they argued—given the challenges in estimating how the magnitude of agglomeration economies varies across regions—that policymakers may have little or no knowledge of how this elasticity varies spatially and hence little basis for preferring one place over another. In addition, if there is no variation in the elasticity across areas, then there will be no aggregate benefits from redistributing activity geographically. Moretti, however, suggested that when there is spatial heterogeneity in the value of this externality, competition among local governments can be an efficient source of place-based policies. He argued that when local governments know the value of the local externality and set locally financed incentives based on it, competition to attract businesses may increase national welfare, despite the potential zero-sum game of attracting businesses to one location rather than another. The reason is that this local policy competition may ultimately arrive at the correct valuation of the externality. Of course, there may be reasons for

¹⁰ Zhang (2007) suggested that venture capital, which might be thought of as supplying capital in a national market, actually tends to favor local industry—looking at the specific context of Silicon Valley.

¹¹ Place-based policies that aim to address a coordination failure and target city size may still only be a second-best response, even in the best-case scenario in which policymakers know the optimal size to capture, for example, the externalities between firms from colocation in the same area. It is more difficult to imagine a policy that fully addresses inefficiencies conditional on location, for example, due to uninternalized increasing returns such as through spillovers, which may depend on the scale or type of the firms' investment.

skepticism about local government motivations and incentives to attract businesses, as other factors—such as the salience of attracting new businesses for winning elections—can easily come into play.

18.2.2 Knowledge spillovers and the knowledge economy

A frequently posited source of agglomeration externalities, which has a long history in urban economics (e.g., [Marshall, 1890](#); [Jacobs, 1961](#)), is knowledge spillovers—or learning in the [Duranton and Puga \(2004\)](#) typology. The hypothesis is that densely populated, diverse urban areas foster experimentation and innovation and facilitate face-to-face interactions that aid the spread of new ideas. More generally, the human capital of others in close proximity can raise everyone's human capital and increase firm productivity, through sharing of knowledge and faster adoption of innovations ([Moretti, 2010](#)). Because knowledge is more likely to spill over from more highly educated workers, due to the knowledge they possess and perhaps the work they do, knowledge spillovers can have more specific predictions than agglomeration economies per se—in particular, that locations more dense in educated workers will be more successful.¹²

Knowledge spillovers can provide a rationale for local policymakers to try to produce or attract skilled workers—for example, through creating or supporting educational institutions, perhaps in particular universities. The public-good characteristics of basic knowledge rationalize public subsidies to research universities in general, but the potential for local knowledge spillovers can rationalize place-based policy. If spillovers increase with geographic proximity and firms are aware of this, then investment in universities may serve to attract innovative firms to the locality. Local governments may take additional steps to increase knowledge spillovers from publicly financed research, such as the creation of business incubators and science parks near research universities or encouraging interactions between universities and businesses, potentially overcoming information or coordination failures.

In addition to the potential value of generating knowledge spillovers from attracting high-skilled workers, [Moretti \(2012\)](#) argued that attracting skilled workers in the knowledge-intensive high-tech sector has large local multipliers relative to other industries. This can occur because of high pay in these jobs, because of demand for business services from this industry, and because high-tech firms appear to attract other high-tech firms. Moreover, there are gains to earnings of others, according to Moretti, from human capital externalities, faster technology adoption, and complementarity with less-skilled labor.

¹² In that sense, knowledge spillovers are viewed as a particular type of agglomeration externality. But the hypothesis has received enough attention in the literature that it merits separate consideration.

Echoing the discussion of agglomeration economies generally, [Glaeser and Gottlieb \(2008\)](#) raised questions about the aggregate welfare implications of policies to try to exploit knowledge spillovers by encouraging moves of educated workers from one place to another or the creation of educated workers in one place rather than another. The argument is similar: There has to be a nonlinear relationship between the density of skilled workers and productivity, so that moving skilled people from one location to another increases productivity more in the target area than it decreases productivity at the origin, and policymakers must know the nature of this relationship. In addition, worker mobility can dissipate the effects of some local policies to exploit human capital spillovers, such as subsidizing education. This may be particularly problematic for higher education as more educated workers are more mobile.¹³

18.2.3 Industry localization

Many of the arguments about matching, sharing, and learning can be applied and may even be more persuasive at the industry level, since the localization of industry employment has been systematically documented for specific tradable sectors for a number of countries. For example, knowledge spillovers may actually have to do more with the presence of workers in the same or related industries, rather than skilled workers, *per se*, in the locality. Other sources of agglomeration economies may be stronger within industries, because the thick labor markets or thick intermediate input markets that may be the engine of agglomeration may operate more within than across industries. Such industry-level externalities may rationalize government policy to try to establish or enhance industry clusters.¹⁴

However, the observation that such industrial clusters exist, and the potential presence of externalities, is not sufficient to justify intervention. For example, for some sectors, access to natural resources rather than the presence of agglomeration economies may drive the location of clusters. [Duranton \(2011\)](#) analyzed the theoretical basis for industrial cluster policies and questioned the magnitude of the returns to clustering for local welfare. His argument that the case for policy is weak rests on the complex nature of the agglomeration externalities and on the costs and benefits of intervention in practice (i.e., on the potential weakness of policy levers if firm and worker mobility is limited and on the evidence of only small-scale effects of clustering on local productivity and wages). Moreover, the optimal size of a cluster, which policy would want to target, is hard to determine in practice and would require knowledge of both the higher benefits

¹³ In this chapter, we do not review the evidence on general education subsidies (see, e.g., [Bound et al., 2004](#); [Bartik, 2009](#)). One might imagine mobility to be less of a problem for education policies that target provision of skills specific to a local industry, such as at the community-college level.

¹⁴ [Bartik \(2003\)](#) and [Matouschek and Robert-Nicoud \(2005\)](#) discussed market imperfections in investment in training that might occur when firms are clustered, noting that in some cases, government involvement in the provision of (industry-specific) training may be justified.

from increasing returns as the cluster grows and the increased costs due to limits on land availability and externalities such as congestion. On the other hand, if a cluster policy were to be pursued, then competition for plants between localities may be more likely to lead to an improved spatial distribution of industry activity, paralleling the argument in [Moretti \(2010\)](#).

18.2.4 Spatial mismatch

Other types of market imperfections, such as frictions in labor or housing markets resulting in involuntary unemployment and reduced mobility, can sometimes rationalize place-based policies, although these may not serve as the standard motivations. One prominent example is the spatial mismatch hypothesis, which—as applied to the United States—argues that the lower employment rate of disadvantaged minorities in urban cores is in part attributable to there being fewer jobs per worker in these areas ([Ihlanfeldt and Sjoquist, 1998](#), p. 851). This can emerge because of the exit of jobs from these areas with the changing industrial structure ([Wilson, 1987](#)) and can persist because of exogenous residential segregation attributable at least in part to discrimination in housing markets.¹⁵

Issues of residential segregation of minorities, immigrants, and the economically disadvantaged generally in areas with diminished job opportunities are hardly limited to the United States, although we know less about this in other countries. Recent studies point to a link between residential segregation and employment or unemployment in France ([Gobillon and Selod, 2007](#)), Belgium ([Dujardin et al., 2008](#)), Sweden ([Åslund et al., 2006](#)), and the United Kingdom ([Fieldhouse, 1999](#)).

The segregation of disadvantaged groups in areas with fewer jobs implies that the wage net of commuting costs for these groups is more likely to be below their reservation wage, so fewer residents of such areas will choose to work, especially among the less-skilled for whom commuting costs represent a larger share of earnings. Customer discrimination against minorities, employer discrimination that deters employers from moving to urban minority areas where wages are lower, and poor information about jobs in other areas ([Ihlanfeldt and Sjoquist, 1998](#)) can reinforce the effects of spatial mismatch.

The gist of the spatial mismatch hypothesis is that the mobility usually assumed in urban economics may be restricted; hence, out-of-equilibrium behavior may persist for a long time. This lack of mobility may undermine some of the concerns expressed in the urban economics literature that place-based policies—often motivated by equity concerns (discussed more below)—can be harmful by inducing poor people to remain in poor areas, if they are likely to remain there anyway.

¹⁵ [Gobillon et al. \(2007\)](#) reviewed theoretical models and hypotheses regarding spatial mismatch.

18.2.5 Network effects

Network effects in labor markets may influence the impact of place-based policies. In network models, employment of some residents increases the flow of information about job opportunities to other residents or the flow of information about workers, reducing search costs and increasing employment (e.g., [Montgomery, 1991](#)). Networks are likely to have a spatial dimension—for example, connecting neighbors. [Hellerstein et al. \(2011, 2014\)](#) and [Bayer et al. \(2008\)](#) reported evidence suggesting that network connections between coresidents (of the same Census tract or even smaller areas) are important. Residence-based labor-market networks can exacerbate the adverse effects of residential segregation on labor-market outcomes for some groups: for example, when social networks are racially (or ethnically) stratified or stratified based on skills.

Network effects do not conventionally arise in discussions of place-based policies. Part of the motivation for a broader perspective that considers this factor is that it may counter some of the criticisms of place-based policies, such as the arguments (discussed more below) that these policies discourage the migration of the disadvantaged to areas with better economic opportunities and that many of the benefits may go to commuters and new residents who have the skills to take advantage of newly created employment opportunities ([Glaeser, 2007](#)).

Coupled with spatial mismatch, network effects may strengthen the case for place-based policies focusing on areas of concentrated disadvantage, because the multipliers that network effects create can amplify the effects of these policies, more so in areas with low employment and perhaps also more so in minority areas where stratification of labor-market networks may imply particularly poor labor-market information.¹⁶ However, even absent the constraints on mobility assumed by the spatial mismatch hypothesis, high concentrations of low-employment areas may help justify policies targeting these areas. For example, one could imagine that in an area with low employment and high crime, utility is not necessarily low enough to induce outward mobility to higher employment, low-crime areas. But that crime surely imposes costs on others, and hence, subsidizing employment of one person to exploit the positive externalities on others' employment (and on crime)—because of networks, for example—can be a prudent policy and more cost-effective in areas with low employment.

This strikes us as a commonsense rationale for place-based policies. Because of crime spillovers between neighborhoods and the location of consumption of urban amenities,

¹⁶ Peer or neighborhood effects can also imply externalities between individuals (see [Topa and Zenou, 2015](#)). For example, the presence of unemployed residents might lead other residents to remain unemployed by changing their norms of behavior ([Wilson, 1987](#)), and, conversely, creating some employment can have virtuous effects on others. Network effects could also diminish the effects of place-based policies. For example, a policy that leads employers to relocate to an area may do little to boost employment opportunities of local residents if the employees of the relocating companies are not networked to local residents.

many city (and suburban) residents—and not only residents of targeted neighborhoods—may be made better off by policies that increase job opportunities in disadvantaged areas. In addition, if network (or peer or neighborhood) effects are important, it may be efficient to target such policies to areas with large concentrations of unemployed people so that the multipliers from these effects can have a greater impact. Viewed this way, network effects may offer a public good that many can take advantage of when some employment opportunities are created. This dovetails with other arguments that place-based policies can in part be justified by the need to correct the underprovision of public goods in poor areas, often because the tax base is insufficient to provide these goods (Crane and Manville, 2008).

18.2.6 Equity motivations for place-based policies

The equity motivation for many place-based policies is to redistribute jobs and income to places where jobs are scarce and incomes are low. Urban economics teaches us that the success of such policies in redistributing jobs and income is complex. It may seem natural, for example, that a state that is concerned with low job opportunities in a specific urban area would try to spur job creation there by using tax or other incentives—such as enterprise zones. However, mobility of people and capital can complicate the effects and potentially undo most or all of the gains from such redistributive policies.

Moretti (2010) developed this argument in some detail. If we think about an enterprise zone type of policy, the subsidies to employment will result in higher wages unless labor supply is infinitely elastic. If labor is mobile, some workers will move to the subsidized area, and as long as housing supply is not infinitely elastic, housing prices and rents will increase, offsetting at least some of the gains to the original residents.¹⁷ Of course, some people in the targeted areas may own property, and for them, the increase in housing prices is a gain. In the extreme case of perfect mobility of labor, utility of each individual is equated across locations both before and after the policy intervention, and the only effect is on land prices that capitalize the place-based subsidy. However, we probably should not consider landowners as the target population for place-based policies based on equity goals.

Thus, other than unlikely knife-edge cases—like infinitely elastic labor supply that implies no wage increases, infinitely elastic housing supply that implies no change in housing prices, or perfect mobility that undoes all gains from place-based policies—mobility probably will partly but not fully undermine the effects of redistributive place-based policies. Nonetheless, the welfare effects can be other than intended. For example, if we rule out perfect mobility of labor and assume that some people have geographic preferences for location, then it is only the marginal workers for whom utility is

¹⁷ Although policy may or may not require workers to live in the area where the subsidies apply, the subsidies will presumably generate some mobility of people into or near those areas.

equated across locations. However, in this case, who gains from the policy may have little to do with the intended effects. Inframarginal workers in the target area gain and those in the other areas (that are taxed) lose, while marginal workers are unaffected. Depending on who these inframarginal workers are, the redistributive effects in terms of welfare may or may not be what policymakers intended. For example, there may be no good reason to believe that the inframarginal workers in the targeted area are the lowest income individuals.

This echoes a broader concern about the targeting of benefits to the disadvantaged via place-based policies. As [Crane and Manville \(2008\)](#) emphasized, given mobility and land-price responses, the jobs created (if they are created) may go to nonpoor residents or migrants, and the gains from land prices seem unlikely to accrue to the poor. At the same time, they suggest that it may be possible (if somewhat utopian) to create institutional arrangements so that the increase in land values is captured by the public and redistributed, to some extent, to the intended beneficiaries. They refer to “Community Benefits Agreements” that specify, for example, that developers who capture the higher land values devote resources to higher wages, affordable housing, social services, etc.

[Glaeser and Gottlieb \(2008\)](#) raised the issue of whether it makes sense to put incentives in place that encourage poor people to stay in poor areas, rather than migrating to places with better economic opportunities. For example, they said, “it is not clear why the federal government spent over \$100 billion after Hurricane Katrina to bring people back to New Orleans, a city that was hardly a beacon of economic opportunity before the storm” (p. 197). This, however, might be an unusual case. If we think, instead, about people living in a poor area who have preferences to stay in that area, then if we could determine that these inframarginal people are the ones we want to help through a place-based policy, one could in principle justify such a policy on equity grounds. Nonetheless, aside from the difficulties of knowing who is and who is not inframarginal, it is not clear that such a policy would be more efficient than subsidizing migration to other areas and perhaps doing more to break down the kinds of barriers to residential mobility emphasized by the spatial mismatch model—if indeed such barriers are important.

18.2.7 Summary and implications for empirical analysis

Two comprehensive reviews of the economics underlying place-based policies, by [Moretti \(2010\)](#) and [Glaeser and Gottlieb \(2008\)](#), disagree to some extent on the efficiency-based rationales for place-based policies, with Moretti taking more a favorable view under some circumstances. In addition, we have suggested some additional efficiency-based arguments that may rationalize place-based policies. But both of these extensive reviews raise serious questions about the equity arguments for place-based policies, with Moretti, for example, concluding that “from the equity point of view, location-based policies aim[ed] at redistributing income from areas with high level of economic activity to areas with low level of economic activity . . . are unlikely to be

effective” (Moretti, 2010, p. 1242). When workers are mobile, it may be better to target people rather than places. It is also important to recognize that equity and efficiency goals in place-based policymaking can end up in conflict. For example, Glaeser and Gottlieb (2008) presented some evidence suggesting that the nonlinearities with regard to knowledge spillovers may be convex, so that subsidizing human capital investment (or in-migration) may be most effective where human capital is already high. Such a policy would tend to increase income disparities between areas.

The preceding discussion highlights some lessons for empirical research even if theory cannot fully pin down a single rationale for the existence of place-based policies. A first test of whether a policy results in welfare gains is whether it generates benefits for the targeted area and, in addition, for the targeted residents. The discussion above points out that the effective incidence of a policy can depend on factors such as the degree of in-migration to an area and the degree of slack in local housing markets. Hence, evaluation should look beyond evidence of effects on local employment to evidence on local unemployment and whether local residents have moved into jobs or whether there have been changes in commuting patterns. We also need to look beyond the effects on local wages, to effects on rents and house prices, to better assess impacts on individuals’ welfare and, further, whether there are heterogeneous effects according to whether people are homeowners or renters or, more generally, by skill or income level. As discussed in Section 18.4, the fact that policy can affect the location incentives of both firms and workers also has practical implications for evaluation methods and the choice of control areas, since displacement can potentially lead to biased estimates of policy effects as well as being of interest as a policy response in its own right.

Two further points emerge that can help guide empirical work. First, local welfare effects might differ substantially from those at an aggregate level. Exploiting agglomeration externalities in one location might come at the expense of (possibly greater) losses of agglomeration benefits in other areas and distortions to the efficient location of economic activity. Second, any local benefits themselves might not be long-lasting. While theory suggests policy could induce a location to shift to a new higher productivity equilibrium, whether this works in practice or whether areas revert to their previous steady state is an important question.

However, it is questionable—based on our own experience with policymakers—that comprehensive welfare statements or calculations carry significant weight in many if not most policy decisions. Rather, policymakers are more likely to start with a goal such as “bring jobs to Detroit.” If we, as urban economists, can simply provide them with rigorous evidence on whether a given policy achieves its stated goal and what other trade-offs—including distributional ones—it entails, we are doing a valuable service and can still help winnow out many policies that do not achieve their goals or have adverse consequences that policymakers do not intend. As a result, most of the rest of this chapter focuses on estimating the causal effects of place-based policies on their targeted outcomes. However, we

touch on evidence on the broader effects of such policies where possible and highlight areas where evidence on the wider welfare implications of these interventions is available.

Even ignoring explicit welfare estimates and calculations, however, there is a potentially significant disconnect between the focus of much empirical research (and, we suspect, policymaker attention) on jobs and the importance of effects on wages for delivering welfare gains to residents of places targeted by place-based policies, because a larger employment response can imply greater deadweight loss from distortions in behavior (Busso et al., 2013). As the theoretical discussion earlier noted, it is when labor is immobile—and hence there is less scope for employment increases in targeted areas—that the welfare gains are more likely to accrue to residents (workers, specifically), rather than property owners. This disconnect may, of course, simply reflect the fact that policymakers place a priority on job creation in specific areas. Alternatively, as Kline and Moretti (2014a) pointed out, when there are labor-market frictions that generate spatial heterogeneity in unemployment, place-based policies like hiring subsidies in certain locations can increase employment (lower unemployment) in the targeted area and increase welfare, so the focus on job creation may be better aligned with effects on welfare.

18.3. EVIDENCE ON THEORETICAL MOTIVATIONS AND BEHAVIORAL HYPOTHESES UNDERLYING PLACE-BASED POLICIES

Glaeser and Gottlieb (2008) and Moretti (2010) provided reviews of evidence regarding the conventional urban economics arguments such as agglomeration economies and knowledge spillovers. We outline some recent evidence briefly but focus on new evidence on the other hypotheses that we believe should be considered in the context of place-based policies, including spatial mismatch and network effects.

18.3.1 Evidence on agglomeration economies

There is now a considerable body of evidence in support of the idea that increased density of economic activity both across and within industries generates positive externalities. Rosenthal and Strange (2004) provided a summary of the evidence on the underlying sources of agglomeration economies. A range of papers have sought to estimate the elasticity of productivity with respect to a measure of the density of employment and generally find elasticities ranging from around 0.01 to 0.10 (see Melo et al., 2009, for a meta-analysis of a wide set of findings). Some recent, but quite distinct, contributions in this area are Combes et al. (2010, 2012) and Greenstone et al. (2010).¹⁸

¹⁸ Much work addresses productivity differences across cities, but some considers agglomeration economies within cities (Rosenthal and Strange, 2003; Fu and Ross, 2013). Some evidence suggests agglomeration economies can attenuate quite rapidly with distance, which is relevant for place-based policies that target small areas.

Combes et al. (2010) carried out a careful analysis addressing identification problems in estimating the relationship between the density of economic activity and productivity. They addressed the issues that a positive relationship between productivity and density may be driven, at least in part, by omitted variables correlated with both the density of employment and productivity, by workers choosing to locate in more productive regions, and by those workers choosing to do so being disproportionately high-skill. Hence, both the quantity and quality of labor are likely to be endogenous. To address the possibility that the quantity of labor may be endogenous, they employed an IV strategy, instrumenting population with historical measures of population density dating back to 1831 and with measures of local geological features including characteristics of the soil and of the terrain, measured by variation in altitude—features that might be expected to have determined where population settlements occurred and how successful they were. To deal with the endogeneity of the quality of labor, they used individual panel data on wages that allowed them to separate location effects from both observed and unobserved worker characteristics. Starting from a benchmark elasticity of around 0.05 between wages and density, they found that controlling for both of these factors led to an estimate of 0.027. In addition, including a measure of market potential (an inverse distance-weighted measure of density across all other areas), to allow for the fact that agglomeration effects may spill across area boundaries, results in their preferred estimate of 0.02, with an elasticity of total factor productivity (TFP) with respect to density of around 0.035.

Combes et al. (2012) examined the extent to which firm selection drives the observed positive relationship between city size and productivity. If competition is increasing in city size, we might expect that low-productivity firms are less likely to survive in larger cities, leading to a positive correlation between city size and average firm productivity, due to greater truncation of the lower tail of the productivity distribution in larger cities. Agglomeration externalities, on the other hand, might be expected to lead to a shift of the observed firm productivity distribution outward as city size increases, as all firms benefit from agglomeration economies, and, if the most productive firms also derive the largest gains, a widening of the distribution at the upper tail.

Their empirical approach estimates the differences in observed firm productivity distributions along these dimensions across more and less dense areas. Their main finding is that selection does not seem to be an important factor in explaining TFP differentials across areas with different employment densities. In addition, they find evidence that firms that are more productive gain more from being in denser environments. Defining denser areas as those with above-median employment density, they find that compared to less dense areas, the productivity gains for firms in the top quartile of the log TFP distribution are approximately 14.4%. In contrast, the gains to firms in the lower quartile from being in denser areas are only 4.8%, implying heterogeneity at the firm level in the degree to which firms might benefit from urbanization externalities. They also find a very similar elasticity of TFP with respect to employment density to Combes et al. (2010), of 0.032.

Greenstone et al. (2010) provided estimates of the magnitude of agglomeration externalities by exploiting a subsidy policy aimed at attracting very large new plants to specific locations in the United States. We discuss the implications of their findings for this category of place-based policy later. Their estimation strategy uses information on runner-up locations as control areas, and their estimates imply that the plant openings resulted in very large productivity spillovers, with TFP in incumbent plants 12% higher than in plants in control areas after 5 years. Of course, as the authors acknowledged, these estimates come from a very specific setting, the opening of a very large new manufacturing plant, for which the winning county may have made the highest bid in anticipation of significant spillover benefits. Effects of this magnitude are therefore not necessarily applicable outside of this policy setting, but are certainly of relevance to the debate about the effects of this type of place-based policy. Greenstone et al. also found evidence of considerable heterogeneity in the magnitude of these externalities both across different locations and across industries. In particular, productivity spillovers are found to be greater in industries that are more similar to the new plant in terms of technologies and human capital requirements, suggesting a role for worker flows between firms and knowledge spillovers (potentially as a result of the former) as sources of agglomeration economies.

Finally, with regard to whether the magnitude of the elasticity of productivity with respect to density varies with the degree of density of economic activity, and hence varies spatially, the evidence described in Section 18.5.5—using an intervention that is perhaps more generalizable (the Tennessee Valley Authority)—does not support the kind of heterogeneity in agglomeration externalities across locations that theory suggests can rationalize place-based policies.

18.3.2 Is there spatial mismatch?

Research testing spatial mismatch in the US context tries to incorporate direct information on access to jobs that is related to either travel time or the extent of jobs (or job growth) nearby (e.g., Ellwood, 1986; Ihlanfeldt and Sjoquist, 1990; Raphael, 1998; Weinberg, 2000). These studies tend to show that blacks live in places with fewer jobs per person and that this lower job access can help explain lower black employment rates, perhaps through the mechanism of blacks facing longer commute times to jobs and hence lower net wages (although Ellwood suggested that the differences may not be large). Evidence of longer commute times for blacks does not necessarily point to spatial mismatch, as simple employment discrimination against blacks can imply fewer job offers and hence on average longer commute times for blacks even if they live in the same places as whites. Overall, two comprehensive reviews argue that there is a good deal of evidence consistent with the spatial mismatch hypothesis (Holzer, 1991; Ihlanfeldt and Sjoquist, 1998), although Jencks and Mayer (1990) provided a more negative assessment of the hypothesis.

Recent work raises questions about the spatial mismatch hypothesis (Hellerstein et al., 2008). In relation to race, the pure spatial mismatch hypothesis implies that it is only the location of jobs, irrespective of whether they are held by blacks or whites (but perhaps conditional on skill), which affects employment prospects. However, if race affects employment—through, for example, discrimination or labor-market networks in which race matters—then even if an area is dense with jobs, black job opportunities may be low. An urban area with large concentrations of black residents, for example, may also be one into which whites tend to commute to work and employers are less likely to hire blacks. In this case, employment problems of low-skilled blacks may not reflect an absence of jobs where they live so much as an absence of jobs available to blacks, which Hellerstein et al. termed as “racial mismatch.”

The authors estimated models for employment including measures of job density not only by location and skill but also by race, using confidential Census information on place of residence and place of work.¹⁹ The evidence is far more consistent with racial mismatch than with simple spatial mismatch. Black job density (the ratio of local jobs held by *blacks* to *black* residents) strongly affects black employment, whereas white job density (the ratio of local jobs held by *whites* to *black* residents) does not. In addition, the own-race relationship is stronger at low skill levels. In a number of specifications, the estimated coefficient on the black job density measure is larger than that of the nonblack or white job density measure by a factor of about 10; the magnitudes are, respectively, about 0.001 and 0.01, with the latter implying that a 10 percentage point increase in black job density raises the employment rate of black men by 1 percentage point. This evidence indicates that for blacks, the spatial distribution of jobs alone is not an important determinant of black urban employment, but rather it is the interaction of the spatial distribution of jobs combined with a racial dimension in hiring, or “racial mismatch,” that matters. In other words, even if blacks reside in areas that are dense in jobs (or dense in jobs at their skill level, as other analyses reveal), if whites tend to hold these jobs, the employment of black residents can be quite low. Reflecting on this, descriptive statistics reported in Hellerstein et al. (2008) show that the density of jobs where blacks live is in fact quite high, even at blacks’ skill levels, suggesting that what is more important is which group is more likely to get hired. And a simple simulation they report showed that if low-skilled blacks were geographically distributed to live where low-skilled whites lived, the black–white employment rate differential would be only marginally smaller (by 0.025, relative to a gap of 0.231). This is precisely because the effect on black employment of white job density—which is the density that would increase most sharply if blacks lived where

¹⁹ These regressions are not plagued by the classic reflection problem that would arise if individual employment were regressed on the local employment rate, because the numerators of the job density measures include both residents and nonresidents (who work but do not live in the area).

whites lived—is so small.²⁰ More recent research establishes that the results are very similar for Hispanics in the US labor market (Hellerstein et al., 2010).²¹

There is evidence for European countries, in the studies cited in Section 18.2.4, which is consistent with spatial mismatch. One of the more compelling studies is Åslund et al. (2006), who studied a refugee settlement policy in Sweden that generates exogenous variation in location, finding that employment rates were lower among those allocated to areas with lower employment rates. However, this evidence typically does not separately consider the density of jobs where people live and the density of jobs for a particular group, as in the racial mismatch analysis. If evidence consistent with spatial mismatch is largely generated by low hiring for minority or ethnic groups, rather than low job availability per se, the case for place-based policies may be weaker than is implied by the spatial mismatch hypothesis. It would therefore be informative to have evidence on spatial versus racial (or ethnic) mismatch for other countries.

18.3.3 Are there important network effects in urban labor markets?

Bayer et al. (2008) presented evidence of labor-market network connections among nearby residents in urban areas. They found that individuals living on the same Census block in Boston are more likely to work on the same Census block than those individuals who do not live on the same block but live in the same block group (a small set of blocks). Because people within block groups are quite homogeneous, their interpretation is that the higher likelihood of working on the same block for those who live on the same block reflects informal labor-market networks based on network connections between those living on the same block (rather than sorting by place of residence and place of work).

Hellerstein et al. (2011) looked instead at whether neighbors work in the same establishment, to test the conjecture that neighborhood labor-market networks operate in part via referrals of neighbors to the employers of those in their network. The method compares the share of an individual's coworkers who are residential neighbors, relative to the share that would result if the establishment hired workers randomly from the geographic areas where *all* individuals who work in the Census tract reside, using matched

²⁰ In a structural model of labor and housing markets focusing on black-white unemployment rate differences in the United States (and African-French differences in France), Gobillon et al. (2013) suggested that spatial factors explain only 10–17.5% of the unemployment rate gap between blacks and whites.

²¹ Andersson et al. (2014) studied the relationship between unemployment duration of workers who experienced mass layoffs and measures of job accessibility, finding that greater job accessibility is associated with shorter durations. The focus on mass layoffs is intended to reduce the correlation between unobserved characteristics of individuals and the accessibility to jobs where they live. The study compares estimates for blacks, for example, using either a general or a race-specific job density measure. The estimated strength of the relationship between accessibility and search duration is similar for both measures. However, it does not estimate a specification including both measures of accessibility simultaneously, as in Hellerstein et al. (2008), without which there is no way to tell whether the race-specific accessibility measure dominates the generic measure.

employer–employee data at the establishment level for the United States (Hellerstein and Neumark, 2003). Labor networks based on the place of residence would imply a higher share of neighbors among a worker’s coworkers than would result from the random hiring process, which in turn simply reflects the likelihood that neighbors tend to work near where they live and hence near other neighbors, irrespective of any connections between them. This difference is normalized by an upper bound for the clustering of neighbors in the same establishment, which arises because, given the size distribution of establishments, perfect sorting by residence-based networks across establishments typically cannot occur.

The evidence indicates that residence-based labor-market networks play an important role in hiring. The “excess clustering” of neighbors in establishments—which is measured as the percent of the maximum systematic sorting of neighbors into the same establishment that could occur that is actually observed—is about 10% for blacks and whites. Controlling for establishment size, this network measure is nearly twice as large for blacks as for whites. Residence-based networks are considerably more important for Hispanics, with the measure rising to 22%, and to around 40% for Hispanic immigrants and those with poor English skills who are less integrated into the labor market and about whom employers may have less reliable information.²²

Labor-market networks that are stratified by race or ethnicity could help explain the racial mismatch evidence and be relevant for place-based policies. Hellerstein et al. (2011) tested for this stratification by constructing the network measure in two different ways: first, treating the relevant set of a black worker’s neighbors and coworkers as including either blacks or whites and hence measuring the extent to which black workers are clustered in establishments with black *or white* neighbors and, second, doing the same computations using only neighbors of the same race. If networks are racially stratified, then the likelihood that a black works with a neighbor regardless of race should be smaller than the likelihood that a black works with a black neighbor—exactly what the evidence suggests. Specifically, the network measure is 40% lower when disregarding the race of neighbors and coworkers, suggesting that labor-market information is less likely to flow between, e.g., black and white coresidents than between coresidents of the same race.

Hellerstein et al. (2010) presented a different kind of analysis, showing that Hispanic job density is most predictive of Hispanic employment in cities in which the Hispanic immigrant population has arrived and grown recently. These are cities in which network contacts may have been especially important in securing employment for new migrants, given that the local economies did not have long histories of Hispanic employment and employers in these areas did not have much experience with Hispanic workers, especially

²² Evidence reported in the paper indicates that the place of residence can be treated as predetermined, potentially influencing place of work, rather than being determined by people who work together choosing to live near each other.

poor English speakers. This study provides further evidence of stratified networks and illustrates how stratified networks can generate evidence of racial or ethnic mismatch.

There is other evidence consistent with ethnically stratified networks. [Kasinitz and Rosenberg \(1996\)](#) studied the Red Hook section of Brooklyn, an area of high unemployment that is populated largely by low-income blacks (and to some extent Hispanics) but with a large number of local jobs in the shipping industry. They found that many local employers hire workers almost exclusively from outside of Red Hook, recruiting employees via social networks within specific (nonblack) ethnic groups. Turning to other countries, [Patacchini and Zenou \(2012\)](#) found that, in the United Kingdom, the probability that one finds a job through social networks is higher if there is a larger share employed among an individual's ethnic group living nearby (accounting for sorting in a couple of ways). [Damm \(2014\)](#), taking advantage of a quasi-experiment involving the settlement of refugee immigrants in Denmark, found that those who were settled in areas with higher overall employment rates of non-Western immigrants and conationals had a greater probability of finding employment and had higher annual earnings if employed.²³

The implications for place-based policies are potentially complex, because racial mismatch or racially stratified networks imply that job creation policies, per se, may do little to help residents in target areas. Effective place-based policies may need to do more to exploit linkages between residents and workers in targeted areas.²⁴

18.4. IDENTIFYING THE EFFECTS OF PLACE-BASED POLICIES

Empirical research on place-based policies focuses, naturally, on estimating the causal effects of these policies on the outcomes of interest. In many respects, the econometric challenges to reliably estimating these effects are similar to the standard program evaluation literature, such as the choice of counterfactuals and the potential endogeneity of where policies are adopted.²⁵ However, there are also a number of issues that are more specific to the analysis of place-based policies. In this section, we discuss these challenges and provide examples of how researchers have addressed them.

²³ Recent research on residential labor market networks using Longitudinal Employer-Household Dynamics (LEHD) data, however, finds less evidence of this kind of ethnic stratification of residence-based labor market networks in the United States ([Hellerstein et al., 2014](#)).

²⁴ [Ananat et al. \(2013\)](#) suggested a potential link between racial mismatch or racially stratified labor market networks and agglomeration economies—presenting evidence that wages rise more with the density of employment in one's industry in the local labor market when measured in terms of workers of the same race and ethnic group.

²⁵ See [Baum-Snow and Ferreira \(2015\)](#) for an overview of identification strategies used to uncover causal effects in urban economics research.

18.4.1 Measuring local areas where policies are implemented and economic outcomes in those areas

One unique challenge is that place-based policies often apply in geographic locations that do not directly map into geographic areas delineated in existing data sources. This issue poses a particular challenge in research on enterprise zones. In California, for example, enterprise zone boundaries do not follow boundaries of Census tracts, zip codes, etc., but are defined by streets and addresses. But because of data availability, tracts or zip codes have often been used to approximate enterprise zone boundaries (e.g., O'Keefe, 2004; Bondonio and Greenbaum, 2007). This introduces measurement error by incorrectly assigning areas (and the workers or businesses in them) as inside or outside the zones (Papke, 1993). Elvery (2009) noted that in California and Florida, if enterprise zones are defined as the areas encompassing all zip codes that overlap with enterprise zones, then the resulting definitions are 6 times larger than the actual zones, and less than half of the population residing in Census tracts that include enterprise zones actually live in enterprise zones. Random incorrect classification of locations creates a bias towards finding no effect of enterprise zones.

Neumark and Kolko (2010), in a study of the California enterprise zones, developed a method of precisely identifying enterprise zone boundaries over time. They start with official lists of street address ranges and the years they were included in the zone and then use GIS software to precisely identify the location of enterprise zones (and appropriate control groups) in each year of their sample.

Once boundaries are defined, data are needed on outcomes of interest within those boundaries and in control areas. Again, this can pose a challenge depending on the geographic information available on workers or firms. Estimating effects for California enterprise zones requires identifying the location of business establishments as inside or outside the zones, because enterprise zone benefits for businesses are based on this location. Neumark and Kolko used a new data source—the National Establishment Time-Series (NETS) (see, e.g., Neumark et al., 2005b)—that provides exact street addresses for establishments in every year. These addresses are then geocoded to obtain precise longitude and latitude, which permits the placement of these establishments in quite exact locations within their enterprise zone (and control area) maps.

18.4.2 Accounting for selective geographic targeting of policies

A second challenge is selecting appropriate control groups for place-based policies. Again, the research on enterprise zones, in which there are three approaches used, is instructive. The first is to identify control areas that are similar to the enterprise zones but where enterprise zone policies did not apply. The second is to use areas that were targeted for enterprise zone designation, but where enterprise zones either were not created or were created at a future date. And the third is to try to deal more explicitly with the endogenous selection of areas for zone designation.

Some studies have used broad control areas where enterprise zone policies did not apply, such as the remaining area of states that are not in enterprise zones (Peters and Fisher, 2002; Lynch and Zax, 2011). However, such broad control areas seem unlikely to provide a valid counterfactual for enterprise zone designation. Others have matched enterprise zone areas to control areas based on the characteristics of the zones or simply nearness to the zone. O’Keefe (2004) and Elvery (2009) matched Census tracts that approximate enterprise zone boundaries to other Census tracts using propensity score matching based on residential and employment characteristics. Of course, propensity score matching does not account for unobservable sources of differences in job growth that may underlie zone designation. None of these studies make use of comparisons of areas observed both before and after enterprise zones were established, while other studies use these matching strategies with before and after comparisons.²⁶

More recent research tries to construct more reliable control groups by using more detailed geographic information on narrow areas. Billings (2009) used a spatial discontinuity model, looking at employment growth in Colorado’s enterprise zones within ¼ mile of the zone boundary and using the area outside the zones within ¼ mile of the zone boundary as the control group.²⁷ Neumark and Kolko (2010) used their detailed GIS maps of the California enterprise zones to pick out a very narrow control ring (1000 ft wide) around the zone, on the presumption that economic conditions, aside from the effects of the enterprise zone, are likely to be very similar in the treated enterprise zone area and the closely surrounding control area. However, nearby and narrow control areas could be subject to displacement effects relative to enterprise zones; this issue is discussed in Section 18.4.4.

Geographic proximity of control areas does not preclude unobserved differences relative to treatment areas, which were the basis of zone designation in the first place. For example, zone areas could have been selected based on responsiveness to zone incentives, in which case the estimation may identify the average treatment effect on the treated (ATT), rather than the average treatment effect (ATE), and the ATT may provide much less reliable guidance to policymakers about the effects of extending the policy to untreated areas. Of course, invalid controls could imply that even the ATT is not identified.

A second approach that may better account for the selection of zones on unobservables is to use as controls geographic areas that were considered or qualified for the treatment or even designated as zones in other periods. For example, Neumark and Kolko (2010) exploited the expansion of zones in their data to compare changes in employment when an area of a zone is designated relative to contemporaneous changes

²⁶ See Papke (1994), Greenbaum and Engberg (2004), and Ham et al. (2011).

²⁷ Freedman (2012) exploited a discontinuity based on poverty eligibility thresholds for tracts.

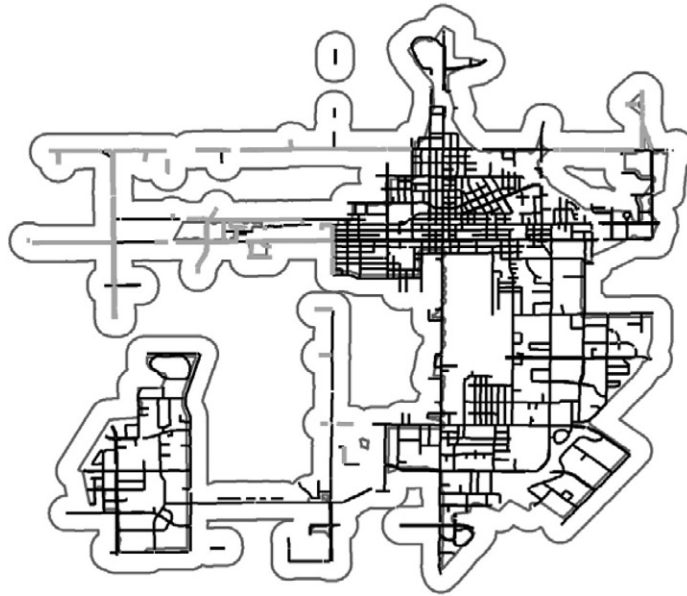


Figure 18.1 Santa Ana Enterprise Zone, initial 1993 designation (thick black lines), 1994 expansion (light gray lines), and control ring (dark gray outer envelope).

in areas that were designated earlier or will be designated later. To illustrate, [Figure 18.1](#) is an example from Neumark and Kolko's study, showing the map for the Santa Ana Enterprise Zone, displaying the initially designated streets, the expansion streets, and the 1000 foot control ring discussed earlier. Identifying effects from comparisons to areas designated at other times can be more reliable than using close areas as controls, because it has been demonstrated through the policy process that the former were appropriate for enterprise zone designation. [Busso et al. \(2013\)](#) used similar strategies, comparing residential employment outcomes in Census tracts that became part of federal Empowerment Zones with outcomes in Census tracts that submitted unsuccessful applications to be designated Empowerment Zones and—paralleling Neumark and Kolko more closely—making comparisons with areas that become parts of zones in the future.

[Hanson \(2009\)](#) also compared employment outcomes in federal Empowerment Zones with unsuccessful applicant areas. However, he also instrumented for zone applicant success based on the political influence of the zone's Congressional representative. The potential advantage of this approach is that nearby control areas or areas that applied for but were not awarded Empowerment Zone status may differ in unmeasured ways that bias the estimated effects. An instrumental variable that predicts which zones succeeded, but does not directly affect the outcomes of interest, mitigates this problem.

18.4.3 Identifying the effects of specific policies when areas are subject to multiple interventions

Place-based policies like enterprise zones may cover areas that are also affected by other geographically targeted policies, sometimes run by different levels of government, and these need to be separated out to estimate the effects of any one policy. Many US cities, for example, have city- or state-designated redevelopment areas that encourage property development to remove urban blight. In California, as an example, hundreds of city and county redevelopment areas overlapped with or were adjacent to enterprise zones in the period covered by the Neumark and Kolko study and hence could affect both treatment and control areas. In addition, the three federal programs—Renewal Communities, Enterprise Communities, and Empowerment Zones—with a variety of benefits similar to those in state enterprise zones, overlapped with state enterprise zones.

To address this problem, the study also used digitized maps of the areas affected by redevelopment policies and federal enterprise zones and incorporated separate identification of these areas into the analysis to isolate the effects of state enterprise zones. Some other studies of enterprise zones pay attention to overlapping federal and state zones, but not redevelopment areas.

18.4.4 Accounting for displacement effects

A potentially serious problem in studying the effects of place-based policies is spillover effects between areas. For example, evidence that enterprise zone designation led to job growth might be regarded quite differently depending on whether the zone created new jobs or employers moved from one area to another to take advantage of enterprise zone credits—which reflects negative spillovers on other areas. Earlier research on the UK enterprise zones found that between 50% and 80% of enterprise zone businesses had relocated into the zones, prompting the British government to phase out the program (Papke, 1993). Of course, relocation does not necessarily imply that a program has not succeeded, because there may have been a number of reasons—reviewed earlier—to try to increase employment in a particular area even at the expense of other areas. Regardless, policymakers should value information on whether job creation in target areas comes at the expense of other areas, or via net job creation.

There can also be positive spillovers. For example, an enterprise zone may increase traffic in a geographic area, spurring demand and hence job growth in nearby areas. In this case, for some research designs, we might find no effect of enterprise zones on employment—or the estimate may simply be biased towards zero—because we are comparing enterprise zones to neighboring areas that were positively affected.

It is difficult to obtain estimates net of spillovers. The usual difference-in-differences approach captures relative effects of a policy on treatment versus control groups, with the assumption that the change over time in the control group was not due to the effect of the

policy. One way to garner evidence on spillover effects is to posit differences across control areas in the likelihood of these effects arising. For example, it seems plausible that positive spillovers are confined to a very narrow geographic area near enterprise zone boundaries. [Neumark and Kolko \(2010\)](#) therefore compared results using a 2500 foot control ring instead of a 1000 foot control ring to see if the estimates of employment effects are stronger using the larger ring in which positive spillovers should be weaker. Similarly, they revert to the 1000 foot control ring but exclude a 100 foot buffer (in any direction) from the enterprise zone boundary. It is less clear, though, that these kinds of approaches are useful in ruling out negative spillovers, since these spillovers may also come from further away, with employers making longer distance moves (although still perhaps within the same city) to take advantage of zone benefits.

18.4.5 Studying the effects of discretionary policies targeting specific firms

Some place-based policies have a discretionary nature—for example, providing subsidies to specific firms to boost investment and employment. Such interventions may be restricted to businesses within targeted geographic areas, but the key characteristic—that not all businesses within the area receive the support—poses an additional identification problem. One question is whether such subsidies are effective in generating additional activity in recipient firms. However, comparisons to nonrecipient firms can be problematic because both the decision to apply for subsidies and the award decision can be endogenous. Those firms that apply may be performing poorly or anticipating a future deterioration in performance, or nonrecipient firms may be judged by government officials to offer less scope for generating additional investment or employment. Hence, nonapplicant and nonrecipient firms will likely have different characteristics than recipient firms, some of them unobservable.

A second question is whether discretionary policies generate benefits external to the recipient firm at the area level. Again, it may be difficult to find suitable controls outside the eligible areas, if those areas in which discretionary subsidies are available have been selected based on specific economic characteristics. In addition, there may be spillovers (positive or negative) from the policy to nonsubsidized firms both within and outside the eligible areas.

[Criscuolo et al. \(2012\)](#) analyzed the effects of a discretionary subsidy policy—Regional Selective Assistance (RSA) in Great Britain. The authors exploited the fact that the set of areas eligible for discretionary subsidies is revised every 7 years according to European Union (EU) state aid rules. Under these rules, subsidies could only be provided in designated areas and then only up to an area-specific maximum subsidy rate. Area eligibility is based on a set of criteria such as GDP per capita, which are measured relative to the EU average. A range of indicators of economic characteristics are used, and hence, areas can change eligibility status due to changes in the prevailing economic conditions in

an area, changes in the indicators used by the European Union to determine eligibility, and changes in economic conditions in other EU member states that will affect the EU average used as a benchmark. The final two of these reasons can be considered exogenous with respect to unobserved characteristics of the areas.

To address the issue that firm eligibility is endogenous with respect to the characteristics of the area in which it is located, the authors used specific features of the eligibility rules as instruments for receipt of an RSA grant. In the estimation, they instrument a posttreatment plant- or firm-level indicator of participation in the program (i.e., grant receipt) with an area-time varying measure of the maximum subsidy rate allowable under EU regulations. They also include plant- or firm-level fixed effects to try to deal with the endogeneity of participation, although this will not deal with problems of time-varying unobservables. This IV strategy likely provides estimates of the ATT, for example in terms of the effects of the subsidy on investment and employment. They also estimate reduced-form specifications, for example regressing log employment at the plant level on the instrument—the maximum subsidy rate at the area level—providing an estimate of the intention-to-treat effect (all plants in an area where the maximum subsidy rate is nonzero being in principle eligible to apply).

The authors also used data aggregated to the area level, for example on employment, to capture any effects due to net entry, in addition to any changes in plant employment at the intensive margin, or due to spillovers across plants within areas. To do this, they regress area-level outcomes on the maximum grant rate determined by the policy rules. They also address the issue of between-area spillovers, for example due to a geographic shift in area eligibility that might lead to displacement of employment to newly eligible areas from contiguous locations, by using a broader geographic aggregation of the data.

Greenstone et al. (2010), as discussed above, provided evidence on the magnitude of agglomeration externalities generated by the opening of a new manufacturing plant. The paper provides a partial evaluation of the benefits of discretionary subsidies offered by local governments in the United States, by examining effects on incumbent plants' TFP growth, net plant entry, and area labor costs. As a novel identification strategy to deal with the endogeneity of the location decision, they exploited information on runner-up localities that narrowly lost out on each plant opening and used these as a counterfactual paired with the winning location. In terms of observed, pre-plant opening trends, the treated and counterfactual sites are highly comparable, much more so than a comparison to all other possible locations. They argued that the use of these near-miss locations as controls should eliminate problems of omitted variables that might otherwise bias comparisons of outcomes across treated and a wide set of nontreated locations.²⁸

²⁸ The identifying assumption is that, conditional on observables, outcomes in the winning and near-miss areas would have evolved identically in the absence of the new plant opening. This rules out other unobserved area-time varying shocks that might differentially affect the paired locations.

For example, if location choices were made based on unobservable characteristics that also positively affect TFP growth, then this form of unobserved heterogeneity across locations would lead to upward-biased estimates of the effect of a new plant opening on this outcome. The use of paired counterfactuals can be considered as a form of one-to-one matching but with the matches determined directly from information on firms' decision-making processes. The authors estimated spillover effects on incumbent plants' TFP in treated counties by estimating plant-level production functions that include dummy variables for each winner–loser county pairing to ensure that the identification of spillovers in the period after plant opening is within each matched pair.

A number of papers evaluate the effects of Law 488, a capital investment subsidy program in Italy, by exploiting a specific feature of the grant allocation process. Applications to the scheme are given a normalized score on the basis of known criteria and then ranked on their score within each region and year. Each region has a preallocated amount of expenditure under the program each year, and hence, projects are funded in rank order until the funding pot is exhausted. These papers exploit the lower ranked, unfunded projects as a control group to address the endogeneity of participation. In a sense, this approach using “near-miss” applicants is analogous to the near-miss locations used by [Greenstone et al. \(2010\)](#) to deal with the endogenous selection of locations.

The fact that unsuccessful projects received a lower ranking means that they differ in their characteristics from the successful applicants. To control for observable characteristics that affect the probability of receiving a subsidy, [Bernini and Pellegrini \(2011\)](#) exploited detailed data on the actual variables used to construct the project-ranking scores. They used this as part of a propensity score matching exercise to control for selection on observables and to ensure common support in observable characteristics across the treatment and control groups. In addition, the authors argued that the fact that the ranking is carried out *within* regions and years and that each region has a different budget for the program in each year generates exogenous variation for a pooled sample of all applicants in the likelihood of being above or below the funding cutoff across these dimensions. Start-up projects are also given priority, and hence for an existing firm making an application, the probability of being funded will also depend on the number of start-ups applying for funding in their host region and year. Hence, it is quite possible for firms with very similar characteristics, and very similar scores, to receive the subsidy in some region-years but not in others.

To control for unobservable time-invariant characteristics, the authors employed a difference-in-differences approach using data on firms in their preapplication year compared with the year after the subsidized project is completed. Clearly, for the control group that did not receive the subsidy, the date at which the project would have been completed needs to be approximated. This is imputed using information on the average completion time, by year, industry, and investment type, from the subsidy recipients.

The authors also argued that spillover effects from subsidized to nonsubsidized firms are unlikely to confound the estimates since subsidized firms make up a very small fraction (around 3%) of manufacturing firms in the eligible regions.

Bronzini and de Blasio (2006) also looked at the effects of Law 488 using a difference-in-differences estimator and using applicants who did not receive a subsidy as controls. Since they found that those firms that score highly and receive a subsidy are a nonrandom sample of all applicants, they tried to address this problem by also adopting an approach akin to a regression discontinuity (RD) design (see Lee and Lemieux, 2010; Baum-Snow and Ferreira, 2015). To do this, they used narrower groups of treated and control firms that are close to the funding cutoff threshold and that have similar scores in the ranking process. These groups are defined as bands, for example firms within plus or minus 30 or 10 percentiles of the ranking distribution of firms around the cutoff threshold. Pelligrini and Muccigrosso (2013) also aimed to identify the impact of Law 488 on the survival of recipient firms using an RD approach. They argued that receipt or nonreceipt of a subsidy close to the budget cutoff point, as in Bernini and Pellegrini (2011), can essentially be considered as random.

18.4.6 Relative versus absolute effects

A final issue is whether empirical research can shed light on aggregate effects of place-based policies and, in particular, whether they result in a zero-sum game, simply relocating activity spatially. Applications of panel data estimators (or other methods of causal inference) can only identify the relative effect of the policy on treated versus control areas, where the latter are by definition assumed to be unaffected by the policy. Hence, such approaches cannot provide information about potential effects of the policy on the control areas, which would let us determine whether the policy had a net positive effect or not across both treated and control areas. Studies that look at displacement or spillover effects (as discussed in Section 18.4.4) can tell us something about impacts on areas not treated directly, but they typically estimate effects for a nearby (often small) area and, to do so, require some other control area that is in turn assumed to be unaffected by the policy.

It possible to make more headway on aggregate effects by relying more on theory. For example, as discussed in Section 18.2, if there are agglomeration externalities that are nonlinear, then moving economic activity can increase aggregate output (assuming activity moves to locations where the externalities are greater). Some evidence on this question comes up in Kline and Moretti's (2014b) evaluation of the Tennessee Valley Authority—a very large-scale place-based policy. As a second example, with enough theoretical structure to estimate welfare effects, one can get evidence on the aggregate effects of a policy (analogous to what we can learn from structural versus nonstructural approaches in other areas of economics). Busso et al.'s (2013) analysis of federal Empowerment Zones presents such an approach and estimates.

18.5. EVIDENCE ON IMPACTS OF POLICY INTERVENTIONS

We now come to the evidence from evaluations of place-based policies. We discuss a variety of types of place-based policies, beginning with enterprise zones. One common theme that emerges across all these types of intervention is that precise policy design matters for the behavioral responses that the policy ultimately delivers, and that some theoretical characterizations of place-based policies as simply setting an optimal city size or delivering a substantial but temporary policy or “big push” that could generate longer run, self-sustaining gains in the presence of agglomeration economies, are far removed from the multifaceted set of incentives that place-based policies provide in practice.

18.5.1 Enterprise zones

The results from earlier studies of enterprise zones varied widely. Many studies failed to find employment effects of enterprise zones, although some of the work (e.g., O’Keefe, 2004, and research reviewed in Wilder and Rubin, 1996) concluded that there are positive employment effects, at least in the short run. Relatively recent overviews of the literature conclude that it is difficult to find evidence of positive employment effects of enterprise zones (Elvery, 2009; Ham et al., 2011; Lynch and Zax, 2011).

However, in the past few years, there have been numerous studies of enterprise zones making creative use of both data and econometric methods to overcome some of the empirical problems involved in evaluating place-based policies in general and enterprise zones in particular. In this section, we discuss this recent research. We begin by discussing the Neumark and Kolko (2010) study of California enterprise zones as an example addressing many of these problems. We then turn to concurrent or more recent evidence, highlighting how other studies address the same research challenges and also trying to resolve what the extensive new literature says and identify the important questions for further research.

As the earlier discussion indicated, the multiple challenges that arise in studying place-based policies imply that the details of the analysis can be quite important. Hence, in some cases, we delve into these details to illustrate the issues that arise and how researchers have addressed these issues and the potential consequences of some of these choices; these are lessons that apply beyond the specific study of enterprise zones. In other cases, the discussion is more cursory and one has to refer back to the original paper for more details.

18.5.1.1 The California enterprise zone program

The California enterprise zone program had multiple goals—not only primarily attracting jobs and businesses and raising employment but also reducing poverty and unemployment and raising incomes in target areas. The program provided a variety of tax incentives to businesses located in designated areas to try to encourage the hiring of economically disadvantaged workers and to spur the creation of businesses. The largest incentive

accounting for the lion's share of the cost was a state tax credit equal to 50% of qualified wages (up to 150% of the minimum wage) in the first year, falling by 10 percentage points each year until reaching zero after 5 years. The main criterion for getting the credit was hiring workers who resided in a Targeted Employment Area (TEA)—a Census tract with low income. However, TEA residents qualified for the hiring credit regardless of the worker's characteristics, and many TEA residents in mixed-income neighborhoods are not disadvantaged. Nevertheless, given that disadvantaged workers earn lower wages, the tax credit could result in a larger relative reduction in the cost of hiring low-skill labor.

Localities applied to the Department of Housing and Community Development to have a geographic area designated as an enterprise zone. Eligibility criteria include job-generating capacity and the level of economic distress measured along a number of dimensions. The area also had to include an industrial or commercial area "contiguous or adjacent to" the distressed area. In addition, the application for enterprise zone status required the preparation of an economic development plan (including marketing, finance and administration of the plan, other local incentives, infrastructure development plans, and information management).²⁹ The hiring credit was paid to firms located in the enterprise zone, but businesses in an enterprise zone could claim hiring credits for employees living in a TEA, which need not be coincident with the enterprise zone. Hence, the program has to be evaluated for businesses located in the zones (or TEA residents), rather than zone residents.

As noted above, Neumark and Kolko exploited the expansion of original zones to construct control areas, while alternative control areas come from very narrow geographic rings around the zone. They defined the original zone and each expansion area, as well as the control rings (when used), as unique "subzones," constructing an observation on each subzone-year pair. They specified regression models for log employment, which include a dummy variable for enterprise zone status and dummy variables for each subzone and year; the year effects account for the possibility that enterprise zones were established in periods of either particularly high- or low-employment growth across all of the regions in the sample. They also included a full set of enterprise zone-year interactions, which allow for an arbitrary pattern of changes over time across the broad area covered by a zone, its expansions, and the associated control ring (when included). Given that the effect of enterprise zone designation is identified off of subzone-level variation, even with these arbitrary changes over time for each enterprise zone, the effect of enterprise zone designation is identified. They also estimated models including subzone-specific linear time trends and models that allow enterprise zone designation to shift the growth rate of employment.

²⁹ The California enterprise zone program was substantially changed in 2013, including eliminating the hiring credit.

Other geographically targeted policies are accounted for in two steps. First, subzone–year pairs are redefined to represent status with regard to not only whether and when they became part of an enterprise zone but also whether and when they became part of a redevelopment area or federal zone, resulting in far more subzones. Second, the specifications are modified to include dummy variables indicating whether each subzone is in a redevelopment area or federal zone in each year.

Across a variety of specifications, there is no evidence that enterprise zones affect employment. The estimates (summarized in [Table 18.2](#), along with the estimates from other studies discussed here) are small, statistically insignificant, and negative as often as they are positive. The statistical power of the evidence is modest, as the confidence intervals for the estimated employment effects are rather large. The baseline model for employment was also estimated with many leads and lags of the enterprise zone dummy variable, to see whether, for example, enterprise zones tended to be established in areas that had transitory downturns in employment relative to other areas, in which case the finding of no effect would be strengthened (because the mean reversion would look like a positive treatment effect). Alternatively, if zones are established in areas doing particularly well just before designation, perhaps because such areas have better organized constituents for capturing an enterprise zone, then the estimated effects from the simple model might fail to detect longer run positive effects of enterprise zone designation on the rate of job growth. Similarly, the many lags allow the data to reveal whether effects of enterprise zones emerge over the longer term. The resulting estimates do not exhibit any evidence of leading or lagged effects, but instead cement the view that enterprise zones in California did not affect employment.

If the enterprise zone program has positive spillovers, encouraging employment growth not only within zone boundaries but also outside zone boundaries, then there might be no evidence of an effect of enterprise zones on employment because enterprise zones are compared to immediately neighboring areas. But the evidence is similar using the larger (2500 foot) control rings. What about negative spillovers, with enterprise zones pulling jobs and businesses away from nearby areas? The similarity of results with and without control rings undermines this possibility. Moreover, such negative spillovers would tend to produce evidence that enterprise zones *do* encourage job growth relative to control areas. Thus, if there were negative spillovers, the conclusion that there are no positive employment effects would only be reinforced. Finally, in the analysis accounting for the overlap between state enterprise zones and redevelopment areas or federal zones, there is similarly no evidence that enterprise zones have positive employment effects, whether or not they are combined with these other local policies.

Enterprise zone programs vary in the level and nature of tax credits and other incentives, as well as in other forms of assistance available to zone businesses—some of which are difficult to quantify and evaluate. This heterogeneity across programs limits how much one can generalize from the study of a single program, and heterogeneous effects

Table 18.2 Summary of evidence on enterprise zones

Study	Country	Program	Results
Neumark and Kolko (2010)	United States	California enterprise zones	No significant evidence of employment effects measured at establishments in zones: estimates range from -1.7% to $+1.8\%$ (levels), with large confidence intervals ($\approx -8\%$ to $+6\%$); no evidence of spillovers
Kolko and Neumark (2010)	United States	California enterprise zones	Zones more involved with marketing and outreach exhibited positive employment effects; zones focused on tax credits exhibited negative effects
Elvery (2009)	United States	California and Florida enterprise zones	No evidence of positive employment effects on zone residents: estimates for California range from -0.4% to -2.6% and for Florida from -1% to -4%
Freedman (2013)	United States	Texas enterprise zone program	Positive effect on employment growth among zone residents ($1-2\%$ per year, sometimes significant); employment effects concentrated in jobs paying less than \$40,000 annually and in construction, manufacturing, retail, and wholesale; positive effects on job growth among zone employers ($3-8\%$ per year, rarely significant) Negative and insignificant effects on share black and with income below the poverty line Significant negative effect on vacancy rate (-4%) Significant positive effect on median home value (10.7%)
Ham et al. (2011)	United States	State enterprise zones, federal Empowerment Zones, federal Enterprise Communities	State programs: significant positive impacts on unemployment rate (-1.6 percentage points), poverty rate (-6.1 percentage points), average wage and salary income ($\approx 1.6\%$), employment ($\approx 3.7\%$) ^a Empowerment Zones: significant positive impacts on unemployment rate (-8.7 percentage points), poverty rate (-8.8 percentage points), average wage and salary income ($\approx 20.6\%$), employment ($\approx 34.2\%$) Enterprise Communities: significant positive impacts on unemployment rate (-2.6 percentage points), poverty rate

Continued

Table 18.2 Summary of evidence on enterprise zones—cont'd

Study	Country	Program	Results
Busso et al. (2013)	United States	Federal Empowerment Zones	(−20.3 percentage points), fraction of households with wage and salary income (4.9 percentage points), average wage and salary income ($\approx 12.7\%$), employment ($\approx 10.7\%$) Positive but insignificant spillovers on neighboring Census tracts Positive and significant effects on job growth in LBD (12–21%), likely concentrated among births and existing establishments with >5 employees Positive and significant effects on employment in Census data (12–19%); magnitudes generally larger for employment in zone of zone residents (15–17%) than nonzone residents (6–16%) Positive generally significant weekly wage effects on zone residents employed in zone (8–13%); magnitudes smaller for zone residents generally (3–5% and usually insignificant) and nonresidents working in zone ($\approx 0\%$) No effects on rents, population, or vacancy rates; large significant positive effects on house values (28–37%)
Reynolds and Rohlin (2014)	United States	Federal Empowerment Zones	Positive effect (1.1%, insignificant) on difference between rent and wage premiums (quality of life) Positive effect (6.4%, significant) on sum of rent and wage premiums (quality of business environment)
Hanson (2009)	United States	Federal Empowerment Zones	OLS estimates: positive significant effect on employment rate (2 percentage points); negative significant effect on poverty rate (−2 percentage points) IV estimates: No effect on employment rate (0 percentage points); insignificant positive effect on poverty rate (2 percentage points)
Hanson and Rohlin (2013)	United States	Federal Empowerment Zones	Negative spillovers on Census tracts that are geographically or “economically” close to zone tracts: negative, generally significant effects on number of establishments (−15.2 to −36.5); negative, sometimes significant effects on employment (−52 to −1223, but many estimates in the range −300 to

Reynolds and Rohlin (2013)	United States	Federal Empowerment Zones	–600); negative spillovers roughly offset the positive effects in directly treated areas Estimates of program effects based on comparison of actual zone tracts to those that are close (using the same definitions) yield positive effects of about the same magnitude as the negative spillover effects Positive significant effects on mean household income (11%), but not on median household income (one-tenth as large) No significant effect on poverty rate (–1 percentage point); significant increase in proportion of households below one-half of poverty line (1.1 percentage points) and in households more than twice the poverty line (1.9 percentage points), coupled with significant reductions in households in between Significant increase in shares of households with income <\$10,000 and above \$100,000 Other results point to higher-skilled, higher-income people moving in: increases in proportions of households with higher education and earnings above \$100,000; increases in housing values for houses valued at \$100,000 or higher, extending above \$300,000 Evidence of spatial variation: increases in proportion of households more than twice the poverty line in areas of zone with above-median poverty rate initially, and increases in proportion below one-half of poverty line in areas of zone with below-median poverty initially
Givord et al. (2013); similar results in Mayer et al. (2012)	France	Zone Franches Urbaines	Positive effects on business creation in and relocation into zone ($\approx 5\text{--}6\%$); similar estimates for employment, but imprecise; positive effects fully offset by negative effects in 300 m ring around zone Briant et al. (2012) reported more positive effects in zones with better access to transportation

^a Approximate percent changes are calculated by dividing their estimates of effects on levels by values in zones reported for 1990.

could help explain why the extensive research literature on the employment effects of enterprise zones is not unanimous in the conclusions it reaches.

Indeed, earlier work provided suggestive evidence of substantial heterogeneity in the effects of enterprise zones, both within and across state enterprise zone programs (Erickson and Friedman, 1990; Elling and Sheldon, 1991; Dowall, 1996), and evidence that enterprise zones were more effective when tax incentives were “complemented by more traditional supports for economic development (e.g., technical assistance, location/site analysis, special staffing)” (Wilder and Rubin, 1996, p. 478). This led Ladd (1994) to suggest that supply-side tax reductions, which are generally uniform across individual zones, are ineffective, whereas interventionist components like technical assistance account for whatever success enterprise zone programs have (p. 202).

In a follow-up study, Kolko and Neumark (2010) explored the associations between the job-creating effects of California enterprise zones and (1) factors relating to the areas in which enterprise zones are established and (2) how enterprise zones are administered. The research used a survey of enterprise zone administrators that asked detailed questions about how active the zone was in using marketing, amending zoning or other local regulations, training workers or operating hiring centers, facilitating earning tax credits, encouraging the building of additional infrastructure, and offering other tax incentives, credits, or discounts on public services at the local level. The analysis uses these responses and information on characteristics of enterprise zone locations that could influence the effectiveness of the program, such as employment density, industry mix, and local demographics.

The estimates point to variation in program effectiveness across individual zones, which has potential implications for features of enterprise zone programs that policymakers and administrators might encourage via both legislation and the selection of sites as enterprise zones. Enterprise zones appear to have a more favorable effect on employment in zones that have a lower share of manufacturing and in zones where managers report doing more marketing and outreach activities. Somewhat surprisingly, a strong focus on helping firms pursue hiring credits made available by the enterprise zone program appears to run counter to job creation efforts, likely because these activities focused more on claiming the tax credits retroactively than on creating jobs currently.³⁰ One implication of these findings is that the overall evidence from the literature on enterprise zones may be too pessimistic and that it may be possible to find ways to make enterprise zones more effective at creating jobs.

Elvery (2009) reached similar conclusions to Neumark and Kolko (2010) for California. He focused on the effects of enterprise zones designated in the mid-1980s on employment of zone residents in the 1986–1990 period. His method matched tracts that are in zones or substantially overlap with them to nonzone tracts using propensity

³⁰ Moreover, until reforms in 2007, “cross-vouchering” was allowed, whereby one zone could collect fees for helping businesses from other zones get credits.

score methods. Elvery also estimated a neighborhood component of employment to capture employment differences across neighborhoods that are not related to the characteristics of individuals in those neighborhoods and also estimated the effect of enterprise zone designation on this neighborhood component, which led to more precision.

Elvery found no evidence of positive effects of enterprise zones on employment—in this case, viewed in terms of employment of residents (rather than employment at enterprise zone establishments as in [Neumark and Kolko \(2010\)](#), which is more appropriate given the distinction between the zones and TEAs). Indeed, his point estimates are always negative, ranging from about -0.4 to -2.6 percentage points (not statistically significant). Elvery did impose some sample restrictions based on age (18–55) and, in some analyses, looked at men only, although it is difficult to imagine that positive effects among the groups he excluded could possibly offset the negative point estimates he reported.

18.5.1.2 Other recent evidence for US state-level and federal programs

A number of recent studies of enterprise zones in the United States address the empirical challenges in different ways and take up different substantive issues, with conclusions that sometimes vary. [Freedman's \(2013\)](#) analysis of the Texas enterprise zone program is a good example of a study that addresses many of the key challenges in the evaluation of enterprise zones. First, the study exploits an unusual feature of the Texas program to construct an appropriate counterfactual, because Census block groups were designated as enterprise zones mechanically, based on whether the poverty rate in the 2000 Census was equal to 20% or greater. Freedman therefore used a regression discontinuity around the 20% cutoff to estimate effects on job growth (his main focus), as well as other outcomes.

Freedman also paid careful attention to estimating the effects for those who should actually be affected by the policy—which is obviously important (as also suggested with respect to California) but not always as simple as it seems. The Texas program has incentives for hiring zone residents, but the firms that hire them do not have to be located in the zone. Rather, employers are designated as “Enterprise Projects” that can claim benefits (sales and use tax refunds of up to \$1.25 million over 5 years) in return for committing to hire a certain share of employees from enterprise zones or who are economically disadvantaged. Freedman suggested that firms are more likely to target workers based on where they live because that is much easier to verify. As a result of these program features, he argued that the hiring effects should be more pronounced for zone residents than for establishments in the zone, and Freedman examined both by using data from the Census Bureau's Longitudinal Employer-Household Dynamics (LEHD) program, which has information on workers' place of residence and place of work. Freedman also accounted for overlapping state and federal enterprise zones and, as discussed below, presented some analyses meant to avoid the effects of spillovers.

Freedman focused most of his analysis on block groups in a narrow band around the 20% poverty threshold, showing that there are no clear breaks around this threshold in

characteristics of block groups other than zone status, which helps to bolster the RD design. The estimates for resident employment indicate a jump in the annual growth rate of employment of 1–2%, which is fairly large when cumulated over a number of years. For workplace employment, the evidence is less clear, consistent with Freedman’s conjecture that the effects should be less evident for employment based on workplace location, although the point estimates are sometimes larger, raising the possibility that something else was occurring in the designated zones. The study also breaks out results by broad industry and broad pay categories, with perhaps the most interesting finding that the employment effects are concentrated in jobs paying less than \$40,000, although because there was not a hiring credit that is a larger share of the pay for low-wage workers, it is not clear why this would necessarily be expected. The paper does not provide any of the usual RD figures for this analysis, however.

Freedman reported analyses excluding control block groups within narrow rings around zones, to avoid possibly overstating the effects because of negative spillovers. Almost none of the estimates are significant for either employment measure when this is done, and in some cases, the positive effects on resident employment become smaller or even negative, consistent with the positive findings being driven more by relocation between nearby areas.

Freedman concluded that “EZ designation has positive effects on resident employment” (p. 340). We are a bit less confident in this conclusion, and the qualification that the effects may largely reflect relocation is important. Nonetheless, this is clearly a very careful study that appropriately addresses the many challenges that arise in drawing causal inferences about the effects of enterprise zones.

Freedman also presented results using the same research design to look at other outcomes, using American Community Survey (ACS) data. Regression estimates indicate a statistically significant 11% increase in median housing values at the 20% poverty threshold, as well as a 4% decline in the share of housing units that are vacant. The point estimates also indicate an increase in population and a decrease in the share black, although these estimates are not close to statistically significant. At the same time, the data indicate no change in median household income. One interpretation—although the evidence can only be taken with a grain of salt given the lack of significant findings—is that enterprise zone designation led to some compositional shifts in the population and that the main effect seems to have been an increase in land value, a finding that arises in other studies, some of which suggest that this may be a principal effect of enterprise zones (e.g., [Hanson, 2009](#)).³¹

³¹ Evidence of the effects of enterprise zones on *commercial* property might be more compelling. [Burnes \(2012\)](#) provided evidence of capitalization of enterprise zone benefits in California into commercial real-estate prices.

Ham et al. (2011) studied state and federal programs and concluded that the programs “have positive, statistically significant, impacts on local labor markets in terms of the unemployment rate, the poverty rate, the fraction with wage and salary income, and employment” (p. 779). Their state-level analysis looks separately at California, Florida, Massachusetts, New York, Ohio, and Oregon, as well as an aggregation of seven other states that have relatively few tracts in zones. As they noted, enterprise zone benefits vary widely across states. For example, the hiring credit—which we would think would be most relevant to job creation—was worth \$35,000 over 5 years in California and was also large in Florida. At the other extreme, Ohio offered \$300 per new employee and Oregon offered no hiring credit. In the federal zones, the main credit was for hiring, with both programs offering a credit of up to \$2400 for 18- to 24-year-olds for the first year of employment and Empowerment Zones also offering a credit of up to \$3000 per employed resident working in the zone for up to 10 years, with the credit declining over time (Busso et al., 2013).

Ham et al.’s econometric approach to the selection problem is to compute a triple-difference estimate. Because they focused only on zones established in the 1990s (or expansions of zones that took place in the same period), their baseline first difference is the difference in outcomes between 2000 and 1990 in the tracts where zones were established. They then subtracted from the 2000–1990 difference the 1990–1980 difference to pick up any differences in linear trends. From the double difference for tracts where zones were established, they then subtracted off the corresponding double difference for three different controls (always using the same years): the nearest tract that was not in a zone, the average of contiguous tracts to the zone that were not in the zone, and the average of all other nonenterprise zone tracts in the state. This estimator can be interpreted as allowing the treatment and control zones to have distinct intercepts and linear trends, but common higher order trends.

To address overlapping programs, Ham et al. restricted attention to tracts affected by only one of the three programs during the 1990s. And to address spillovers, they estimated treatment effects for the nearest nonenterprise zone tract to each enterprise zone tract, using the second-nearest nonenterprise zone tract as the comparison. A potential downside of this approach is that the comparison tract (or tracts) for the actual enterprise zone tract and the tract where there are potentially spillovers are not the same—in contrast to comparing both to a more distant tract that clearly was not affected by enterprise zone designation.

The results Ham et al. reported for the combined (average) effect of state enterprise zones and for the two types of federal zones are almost always strong and positive. As summarized in Table 18.2, across the three types of zones, the authors generally found positive and significant effects on the outcomes they considered. Moreover, they concluded that spillover effects are if anything positive, although in general the evidence is statistically weak and sometimes points in the other direction; certainly, there is no strong indication of negative spillovers. At the same time, some features of their estimates are surprising and hard to interpret. First, looking at the estimates for the federal zones and

for the state zones averaging across all the states they considered (which is what [Table 18.2](#) reports), the estimates are often implausibly large, such as an increase in employment of around 34% from federal Empowerment Zones and a reduction in the poverty rate of 20.3 percentage points from federal Enterprise Communities. More generally, the positive effects they estimate for Enterprise Communities are surprising, given that other researchers regard Enterprise Community benefits as inconsequential relative to Empowerment Zone benefits ([Busso et al., 2013](#); [Hanson and Rohlin, 2013](#)).

Second, the cross-state variation in estimated effects is surprising. The estimated employment effect for California is small and negative, whereas only for Ohio is there a significant positive effect. Yet California had a huge hiring credit, whereas Ohio's was only \$300. And Oregon, which has the second-largest point estimate for the employment effect, had no hiring credit. They do estimate a large employment effect for Florida (not statistically significant), and Florida has a large hiring credit; yet [Elvery's \(2009\)](#) estimates for Florida for the previous decade are consistently negative.³² To increase precision, Ham et al. presented estimates pooling the data across states. But the large policy differences across state enterprise zone programs make this questionable. If one accepted the constrained estimate, one would be equally likely to believe it applied with no hiring credit or a \$35,000 hiring credit.

These kinds of findings indicate that it would be useful for future research to try to tie the effects of enterprise zone policies to features of those policies and exploit variation in the generosity of the policy to a greater degree in evaluation. This could include information on the value of hiring credits and other features that would make such policies more likely to lead to employment increases, such as requiring that the number of employees grow in order for firms to qualify for incentives. One possibility, although it cannot reconcile all the conflicting findings, is that the effects of enterprise zones arise from something other than hiring credits; this issue resurfaces in research focusing on federal Empowerment Zones, to which we turn next.

[Busso et al. \(2013\)](#) studied the effect of federal Empowerment Zones. They compared outcomes in the six urban communities that were awarded Empowerment Zones with the full range of benefits and credits, to matched tracts of rejected zone applicant areas and areas in future zones. The comparison to future zones has parallels to [Neumark and Kolko \(2010\)](#), but differs from Ham et al. in using tracts from other cities and states as controls, rather than nearby tracts.³³ In addition to substantial hiring credits, Empowerment Zones

³² [Billings \(2009\)](#) reported positive employment effects for new establishments in Colorado's enterprise zones, based on a border discontinuity design, looking at 1990–2000 data, even though Colorado's hiring credit is minor (only \$500, increasing to \$1000 under some circumstances, plus a job training credit).

³³ Busso et al. did not address overlap between federal and state enterprise zone programs. They also argued that spillovers are unlikely to affect their estimates because most rejected and future zones are in different cities. But that does not rule out their estimated effects reflecting the relocation of businesses into a zone from nearby.

received block grants of up to \$100 million for purposes such as business assistance, infrastructure investment, and training programs. For nearly all of the cities in which zones were rejected, Enterprise Community status was awarded instead; these areas did not receive major block grants and had much more restricted hiring credits.³⁴ Busso et al. used 1980, 1990, and 2000 Census data including confidential information on where people live and where they work, by tract, and establishment-level data from the Longitudinal Business Database (LBD) from 1987, 1992, and 2000. In both cases, they focused on the estimated impact of Empowerment Zones designated in 1993 on changes over the 1990s.

Busso et al. found, in both the Census and LBD data, that Empowerment Zone designation appears to generate substantial job growth—around 21.3% in the LBD and 12.2% (not significant) in the Census data. Moreover, the Census data suggest that there were increases in jobs in the zone held by residents (17.6%), but less evidence of such effects for nonresidents (6.4%, not significant). The Census data also point to large increases in nonzone employment of zone residents (12.3%, not significant), which suggests that there may have been effects due to factors other than zone incentives.

They also found evidence of positive wage effects on zone residents working in the zone (12%), but no effect on wages of nonresidents working in the zone. The positive point estimate for the employment of nonresidents in the zone is consistent with zone designation increasing the productivity of labor in the zone, but we should not necessarily expect a positive wage effect because the hiring credit should not affect wages of nonzone residents relative to what they earn elsewhere (in equilibrium).

The fairly large estimated employment effects for zone residents working outside the zone suggest that the effects on zone employment that Busso et al. estimated are not fully attributable to the hiring credit. And some of the other estimates they reported (in their Table 5) of employment effects on nonzone residents—although not their preferred estimates—are quite large. The block grants are large, and there is some evidence—although Busso et al. appropriately noted that it is far from rigorous—that the block grants (or something else about the zones) may have attracted large amounts of outside private capital. This could have boosted employment in the zone of nonresidents and, perhaps through spillovers, employment of zone residents outside the zone. If in fact the block grants played a major role, this may help square the results of Busso et al. with those of [Neumark and Kolko \(2010\)](#), who found no effect of the California enterprise zone program that very much concentrated its incentives on hiring credits.

³⁴ Note that these cities are included as controls (or potential controls, subject to trimming based on propensity score matching) by Busso et al., whereas Ham et al. studied these as a separate group of treated cities and sometimes found larger effects than for Empowerment Zones. Given the potentially large benefits suggested by the Busso et al. results, this may further reduce the plausibility of the results for Enterprise Communities found by Ham et al.

The authors also developed a stylized general equilibrium model that captures the welfare implications discussed in [Section 18.2](#)—that the welfare gains from place-based policies depend on whether many agents are inframarginal in their preferences over places to live and work, which is more likely when their preferences are more heterogeneous. In their modeling framework, one can express the effects on welfare in terms of elasticities of the various marginal responses one can estimate from the data (some more precisely or convincingly than others). Adding estimated effects on housing prices, rents, and population, their welfare calculations point to potentially large gains—in large part because an absence of positive effects of zone designation on population or rents suggests that large migration responses do not dissipate the gains of the program. Despite finding no effects on rents, Busso et al. found large effects on house prices. They suspected these are inflated, in part because they did not find an effect on vacancies or on out-migration, but other estimates still point to positive effects. And they suggested that rents may be sticky in the short run because of rent control. However, other evaluations of the Empowerment Zones discussed below also report evidence of house price increases, as well as some evidence of compositional shifts towards the more skilled and a decline in vacancies, suggesting smaller welfare gains.

The authors estimated relatively little deadweight loss from the program, based on calculations and auxiliary information used to estimate jobs created in the targeted areas, although they noted that it is difficult to pin down estimates of the number of jobs created in the zones, for zone workers, in the covered sector. They estimated substantial welfare gains. The bulk of the estimated welfare gains come from positive effects on house prices in the zones, raising the question of whether the program is achieving its distributional goals. A key question is whether the block grants raised productivity of workers in the zone. Under fairly small gains (0.5%), the benefits would outweigh the costs, but we do not know the actual impact of these block grants. Another question is whether the evidence of positive employment effects is attributable to the effects of these block grants. It would therefore be useful to know how these block grants were used and whether there is corroborating evidence that they had effects that boosted employment—perhaps in conjunction with zone benefits—in light of other evidence that even generous enterprise zone hiring credits do not increase employment. This discussion again emphasizes that it would be useful, in future research, to try to parse out the effects of different dimensions of enterprise zone policies.

[Reynolds and Rohlin \(2014\)](#) approached the problem of evaluating the welfare consequences of these zones in a different way, by embedding the analysis in the quality of life framework standard to urban economics, estimating the effects of Empowerment Zone designation on wages and rents in hedonic equations. They found notable increases in what is termed the “quality of the business environment,” which is captured in the summed effects on rents and wages. And they found modest increases in the quality of life for individuals, captured in the difference between rents and wages (what

individuals are willing to pay to live in an area).³⁵ Moreover, Reynolds and Rohlin presented evidence suggesting that the role of the wage credit is negligible. This is potentially important because the value of the wage credit only lasts as long as the wage credit is available, whereas other sources of gains may be more permanent, such as agglomeration effects from firms moving into the area.

The impact of federal Empowerment Zones has been considered in a number of other studies, most of which are less favorable than the Busso et al. evaluation on many dimensions. [Hanson \(2009\)](#) extended the analysis of Empowerment Zones by considering endogenous selection from among the applicants, which can induce selection on unobservables that is not addressed by matching methods. If selection is based on unobserved improvements in economic conditions, there is a bias towards finding positive effects. The results cannot be directly compared to those of Busso et al. because there are other differences in both data and research design. It would likely be highly informative to explore the issues discussed in the various papers assessing the federal program in a single study that held the data constant and focused solely on the substantive issues addressed by different research designs.

Hanson instruments for zone designation using representation of the areas encompassing the proposed zones on the powerful US House Committee on Ways and Means. Hanson considered possible reasons the instrument may be invalid, such as if representation on the Committee Ways on Means yields other economic benefits (or costs) to the same districts. He presented some evidence suggesting this is not the case, although that begs the question of why members exert power over zone designation but not other public resources and, if members do exert other influences on economic outcomes, the IV would be invalid. The estimates without instrumenting indicate that Empowerment Zone designation increased employment significantly, by 2 percentage points, and reduced poverty significantly, also by 2 percentage points. However, the IV estimates indicate no effect on employment and a positive but insignificant effect on poverty.³⁶ Hanson concluded that “OLS specifications over-estimate the effect of the EZ program on increasing resident employment and decreasing poverty” (p. 728).

[Hanson and Rohlin \(2013\)](#) attempted to directly estimate the spillover effects of federal Empowerment Zones on nearby or similar areas—effects that could be negative

³⁵ This approach may not capture welfare in the sense of the aggregation of utility across individuals. But that is a very tall order and many economists are skeptical of the stylized models needed to engage in such calculations. However, by capturing these two dimensions of the effects of enterprise zones, Reynolds and Rohlin successfully encompassed most of what other research talks about as the criteria for deciding whether these policies are effective.

³⁶ Hanson also finds that OLS estimates of the effects on median residential property values indicate insignificant increases of around \$6600, which is smaller than reported by Busso et al. or by [Reynolds and Rohlin \(2013\)](#), discussed below. And his IV estimates are huge, indicating implausible increases of over \$100,000.

or positive. They identified tracts that are close to the Empowerment Zones, based on either geography or economic similarity, and compared changes from before and after zone designation for the close tracts to what happened in tracts that were close—on the same measure—to the rejected applicants in other cities (which became Enterprise Communities). The evidence points to negative spillover effects. For establishment counts, the estimated effect ranges from -15.2 to -36.5 , with almost all of the estimates statistically significant. The employment effects are more variable and statistically significant about half the time; but many are in the range of -300 to -600 .³⁷ Moreover, when they estimated program effects based on comparisons of the actual zone tracts to those that are close (using the same definitions), the positive effects are of about the same magnitude in absolute value as the negative spillover effects, suggesting that Empowerment Zones are, to a first-order approximation, simply creating relocation of economic activity.

A recent study by [Reynolds and Rohlin \(2013\)](#) instead emphasizes that evidence of positive mean effects of Empowerment Zones masks distributional effects that are much less favorable to the disadvantaged. Using similar but not identical methods and data, their results indicated that the zones were advantageous to high-skilled, high-income people who to some extent likely moved into Empowerment Zones because the program made these areas more attractive and were neutral or even harmful to the impoverished residents of these zones. They largely replicate the Busso et al. findings that Empowerment Zone designation boosted mean wages and employment—although in their case, this is documented in terms of mean annual household income, which they estimate rose by around 11%. However, the effect on median household income was only \$250, and not statistically significant, and Empowerment Zone designation had no detectable impact on the poverty rate, with an insignificant 1% decline.³⁸

Moreover, they find increases in the proportion of households below one-half the poverty line, commonly termed “extreme poverty,” with an estimated increase of 1.1 percentage points. When they look at effects across bins of the household income distribution, the only sizable (and significant) increase occurs for households earning at least \$100,000, which is unlikely to be directly attributable to Empowerment Zone incentives (since the hiring credit represents a much larger percentage of pay for low-wage workers), and an increase in the share of households with an income of less than \$10,000.³⁹ They also present evidence of increases in the share of people with higher

³⁷ The authors do not report the means needed to calculate percentage effects, but these are large numbers relative to tract population, which averages about 4000.

³⁸ Their data are closer to [Hanson \(2009\)](#), although their approach is closer to Busso et al. The poverty estimate is not that different from Hanson, who found a 2 percentage point decline.

³⁹ They have to use income categories that are fixed in nominal terms across the years. Clearly, then, there will be some upward drift into higher income bins from inflation. However, the authors argue that since this occurs in both the Empowerment Zone and non-Empowerment Zone cities, it does not affect the triple-difference estimates.

education (some college or more), most likely consistent with inflows of higher skilled people into the areas designated as Empowerment Zones—like the increased share with high incomes. Finally, when they break up the zones into tracts with initially above-versus below-median poverty rates, they find that the positive income effects (at \$100,000 or above) occur solely in the lower poverty tracts, whereas there is evidence (not quite statistically significant) that the increase in the share of households with less than \$10,000 in income occurs in the high-poverty tracts.⁴⁰

It is hard to be definitive about what reflects an impact on residents versus a composition effect. But it is unlikely that lower income households would have replaced higher income households in response to Empowerment Zone designation. In contrast, the increase in education levels (and higher incomes) seems likely to be a composition effect. Thus, these results present a more negative portrait of federal Empowerment Zones as failing to deliver on the goal of helping low-income families in these areas and make an important addition to the growing literature arguing that we need to assess the distributional effects of public policies—especially those intended to influence the distribution of income (e.g., [Neumark et al., 2005a](#); [Bitler et al., 2006](#)). The apparently adverse distributional effects do not necessarily contradict Busso et al.’s estimates of the value created by the Empowerment Zone program. But they certainly raise questions about how to evaluate the gains from a social welfare perspective. Moreover, if the zones generated housing price gains from gentrification, it is entirely possible that there were offsetting housing price effects in areas that were unlikely to be weighted heavily as controls for the tracts designated as Empowerment Zones, as there is no reason to think that the higher skilled in-migrants came from other disadvantaged areas.

An obvious question is why Empowerment Zone designation would have mainly benefitted higher income households, including perhaps enticing some to move into the zones. One possibility is that the block grants were spent on things that had this gentrification effect, rather than on activities or programs that increased job opportunities for low-income zone residents. There does not seem to be a lot of information on how these funds were used, although a 2006 GAO study ([US Government Accountability Office, 2006](#)) gives brief summaries of what each of the zones did, noting that some Empowerment Zones and Enterprise Communities focused more on community development than economic opportunity. It also cites some specific examples that could be viewed as having these types of effects, such as contributing financial assistance to a 275,000

⁴⁰ The authors’ conclusions differ from those of [Freedman \(2013\)](#), who suggests that “Texas’ EZ Program had a positive effect on communities, but one that was largely confined to households in the lower end of the income distribution” (p. 340). However, this is not based on as comprehensive a distributional analysis as Reynolds and Rohlin’s, but rather seems to derive from evidence of the positive effects discussed earlier, coupled with no effect on median income in the ACS data.

square foot retail complex in Harlem, housing development in Detroit, and cleaning up vacant lots in Philadelphia.

Whether one shares the positive assessment of Empowerment Zones of [Busso et al. \(2013\)](#) or the more negative assessment of [Reynolds and Rohlin \(2013\)](#), it would be useful to learn more about how the large amount of block grant funds was spent, since much other evidence on enterprise zones suggests that the other components of the policy, like hiring credits, had little impact. Of course, the Busso et al. and Reynolds and Rohlin studies focused on only six Empowerment Zones, so it is difficult to imagine making much headway based on differences among these zones and suggesting also that we should be wary of generalizing either set of results. However, the program was expanded to an additional 23 Empowerment Zones (21 new ones plus two that were upgraded) beyond the original six that these two papers study, so there may be more that can be done with the additional zones, although much smaller discretionary sums, rather than large block grants, were made available to these additional zones ([Mullock, 2002](#); [US Government Accountability Office, 2006](#)).

Another question is whether enterprise zone hiring credits could be made more effective depending on how they are structured. For example, the hiring credits under the California enterprise zone rules did not do anything to require or verify job growth. In contrast, Ohio's program requires employment growth and that employment was not reduced or a facility closed outside the enterprise zone program to facilitate growth in the zone (although one might wonder how this is determined). It is conceivable that these design features matter. For example, in related work on hiring credits generally, which specifically require job or payroll growth, [Neumark and Grijalva \(2013\)](#) found that allowing states to claw back credits if the job creation goals are not met appears to make these credits more effective. This kind of clawback feature, with payment linked to specified job or investment targets, is used in some of the discretionary subsidy policies discussed later.

18.5.1.3 Evidence from other countries

The French Zones Franches Urbaines (ZFU's) are enterprise zones modeled similarly to US programs. Firms with fewer than 50 employees located within the zones' boundaries are exempt, for 5 years, from local business, corporate, and property taxes, as well as social security contributions on the fraction of salaries lower than 1.4 times the minimum wage; the exemptions are then phased out slowly after the initial 5 years. This program therefore appears remarkably generous and is simple in that it is based solely on location. One might expect that, given that the incentives are tied so strongly to location, the program would have the most impact on births of businesses in the zones or relocations of businesses into the zones.

In studying ZFU's, [Givord et al. \(2013\)](#) addressed the key challenges that arise in evaluating place-based policies. First, the paper uses rich administrative panel data on establishments with information on precise location, the creation date of the business at that

location (and whether it was a birth or a relocation), and the date of bankruptcy if that occurred; the precise location information is important because enterprise zone boundaries do not coincide with existing jurisdictional boundaries.⁴¹ The authors had administrative data on employment and salaries, fiscal records of taxes paid, and data on other business outcomes.

Second, the paper uses compelling control groups. The implementation of the ZFU program occurred in three phases. In the first, in 1977, the government defined 44 areas that were granted ZFU status. An additional 416 areas were identified as slightly less distressed economically. Then, in 2004, 41 new ZFUs were created, all chosen from the second group of 416. The authors' identification strategy compares changes in these 41 new ZFUs with changes in some of the remaining groups, known as ZRUs, which were initially identified as distressed, but not designated as ZFUs in the second round.⁴² The zones designated in the second round might be expected to be relatively more distressed. But the committee selecting these zones had to follow precise guidelines based on an index calculated from population, unemployment rates, tax revenues, the proportion of youths, and the proportion of dropouts. Given that the authors can construct the same index, they are able to compare the ZFUs designated in the second round with zones having the same index value that were not designated and show that the main selection criterion was that new ZFUs were far from those established in the first round (presumably to achieve a more even geographic distribution) and near to other ZRUs (allowing mergers to achieve the minimum population size). They use propensity score matching to candidates among the ZRUs to estimate treatment effects, matching on the variables used in the index, these distance measures, and other variables including outcomes prior to ZFU designation.⁴³

Third, the paper pays close attention to negative spillovers, which we might expect to be particularly important in the case of a program that simply pays incentives based on location. Indeed, this study grapples more directly with relocation than the other recent

⁴¹ Givord et al. focused on single-establishment firms because the tax information is at the firm level, and other data supposed to be at the establishment level are sometimes aggregated across establishments. Since ZFU incentives apply to firms with fewer than 50 employees, focusing on single-establishment firms should capture most affected businesses. The discussion below refers to "businesses," but the meaning is single-establishment firms.

⁴² These latter areas, called Zones de Redynamisation Urbaine (ZRUs), also have some tax incentives, but these were negligible during the study period.

⁴³ In a closely related paper on the same program, using the same data source, [Mayer et al. \(2012\)](#) compared ZFU and non-ZFU areas in the same city. As they argued, this lets them control for city-specific differences in policy, transportation infrastructure, etc. The findings in Mayer et al. are quite similar on many dimensions, both in terms of responsiveness to zone incentives, and the finding in Givord et al. that the positive effects in zones are largely offset by diversion from areas near the zone (in the case of Mayer et al., the rest of the city).

studies. Relocations are of interest as a direct manifestation of the kinds of negative spillovers discussed earlier.

Givord et al. first showed evidence not usually available—that the ZFU program resulted in substantially lower business taxes and social security contributions paid relative to the control areas. (One potential explanation for an absence of employment effects in many US studies is that most companies do not take advantage of the tax credits.) Turning to outcomes, the ZFU program positively affected the number of businesses located in the treated areas, via both births and relocations. The program boosted the number of establishments by about 5–7%. This is due roughly equally to births and relocations, but compared to average relocation versus birth rates, the relative impact on relocations is much higher—about 100% compared to 25%.

When the authors split the sample into existing businesses, there is no evidence of an employment effect; the point estimates range from –6% to 9% and are never significant. Thus, it seems likely that if there was employment growth, it came from the new businesses that moved into the ZFUs, although with the imprecise estimates we cannot be sure. Mayer et al. (2012), in their analysis of the same program, reported results indicating that only establishments with 50 or fewer workers (and with sales below the maximum to be eligible) were affected by the policy, in terms of the decision to locate in a ZFU. (They do not study employment responses among businesses already in the ZFU.)

Finally, given the evidence suggesting that the response is strongest on the relocation margin, Givord et al. explored negative spillovers on nearby areas. These spillovers could also be manifested in business births, because the incentives can influence the decision about where to open a new business. Using a procedure very similar to that of Neumark and Kolko (2010), the authors constructed 300 m rings around their ZFUs and estimate the impact of the ZFU incentives on these areas. Their comparisons are based on similar rings around their control areas (the ZRUs). Interestingly, the results mirror quite closely—albeit with opposite sign—the effects on activity *inside* the ZFUs. Fewer businesses are created (through both the birth and relocation channel, although only the estimates for the latter are statistically significant). And the employment estimates are generally negative and of similar magnitude (in absolute value) to the positive estimates inside the ZFUs, although again insignificant.⁴⁴ Although there could be gains from relocation, it seems hard to make the case that there are gains from reallocation of activity over such small geographic areas or that such gains could offset the substantial foregone tax revenue that the study documents.

⁴⁴ Mayer et al. (2012) also concluded that the ZFU program mainly led to diversion, with even stronger evidence that the response is centered on the relocation margin. In their analysis, this is based on evidence that the overall flow of establishments into municipalities was not affected by the designation of a ZFU within the municipality.

A recent paper by Briant et al. (2012) extends the analysis of ZFUs to consider heterogeneity in the effects of the program based on geographic features, including access to transportation and barriers and distance between targeted areas and main employment centers. Briant et al. used a simple difference-in-differences strategy based on geographic subunits of municipalities. The effects of ZFU designation vary in ways we might expect with geographic features of the treated areas. In particular, better access to transportation by roads or trains is associated with a larger positive effect in attracting firms. Moreover, some of these geographic features appear to do more to boost the creation of new firms rather than just relocations of existing firms, in contrast to the results for homogeneous treatment effects in Givord et al., which suggested that relocation was the primary response. The paper clearly presents some intriguing findings that bear further study in the context of enterprise zones in France, as well as other countries, and emphasizes the point made with respect to the US literature as well—that we need to learn more about when enterprise zones are more effective or less effective.

18.5.1.4 Summary of evidence on enterprise zones

The research literature on enterprise zones is noteworthy for two reasons. First, it provides a fairly comprehensive examination of the effects of this particular place-based policy. Second, it is illustrative of the kinds of analyses that need to be undertaken to evaluate place-based policies generally, exhibiting careful and creative efforts to address the numerous challenges that arise in evaluating the effects of these programs and—especially in the most recent studies—attention to issues beyond mean effects, which are needed to evaluate the distributional and perhaps even the welfare effects of these policies. In this summary section, we do not reiterate the general research contributions of this literature, which have been highlighted in the earlier discussion. But we do attempt to summarize the findings for enterprise zones specifically.

Research on three specific state programs (California, Florida, and Texas) concludes that two generate no employment effects and the third (on Texas) finds positive effects concentrated on lower pay jobs. One study looking at numerous states also finds some positive employment effects, but they do not appear to be tied in any way to hiring credits. Thus, evidence on whether these state programs created jobs is mixed, while a stronger case can be made that if they did create jobs, it was not because of the hiring credits highlighted in many state enterprise zone programs.

Evidence from analyses of the US federal Empowerment Zones program is also mixed. One study finds strong effects on job growth and wages, whereas another suggests that if we account for endogenous selection of zones, there is no evidence of beneficial effects. Moreover, if there are benefits, they appear to accrue to higher income households. If one concludes that the federal program was beneficial, it seems plausible that the large block grants associated with Empowerment Zones played an important role, although verifying that will be challenging given the small number of affected zones,

and these grants may have done more to increase the attractiveness of zones to higher income people. The later round Empowerment Zones did not receive these large block grants. Thus, comparison of the effects of the later round Empowerment Zones with the first-round zones could be informative about these questions.

The evidence on spillovers is also mixed for the United States, with some studies suggesting negative spillovers that offset program benefits. There is evidence of strong spillovers in France, where zones offering significant tax breaks resulted mainly in the relocation of economic activity from nearby areas to inside the zones. There may be reasons policymakers want to relocate economic activity to some areas even if this is solely at the expense of other areas. But clearly the case for place-based policies is harder to make if this is what happens, especially for relocation over small areas. At a minimum, we would presumably want to see evidence of other beneficial effects of this relocation of economic activity. Indeed reflecting the preference for boosting firm births, for example, rather than relocations, some states have inserted provisions that bar relocating businesses from obtaining enterprise zone benefits (Wilder and Rubin (1996) and the current Ohio program).⁴⁵

Finally, theoretical modeling of the effects of place-based policies indicates that the welfare implications depend also on the effects on housing prices and migration responses. On these issues, too, the conclusions are mixed, with Busso et al. (2013) staking out a position that rents do not increase and that there are no compositional shifts, leading to a rather strong positive evaluation, while other research emphasizes housing price increases (which may not only reflect welfare gains but also have distributional consequences), compositional shifts, and gains for higher income households to a much greater extent.

What do we make of all this? We draw two main conclusions. First, it is very hard to make the case that the research establishes the effectiveness of enterprise zones in terms of job creation or welfare gains, although there clearly are some studies pointing to positive effects. Further progress requires effort to figure out what features of these programs can make them more effective, following on some early efforts in this direction in the existing research. Second, although there have been a slew of new studies in the past few years—and even many studies focusing on the same program—there has not been enough of an attempt to reconcile the disparate evidence. This kind of careful, often painstaking, work may well help sharpen the conclusions from a research literature in which the findings remain rather at odds.

18.5.2 Place-based policies that account for network effects⁴⁶

Earlier, we discussed how place-based policies might be more effective if they took advantage of labor-market networks involving the residents of targeted locations. Ladd (1994) echoed this point, suggesting that place-based policies should recognize

⁴⁵ See http://development.ohio.gov/bs/bs_oezp.htm (viewed 6 September 2013).

⁴⁶ The discussion in this subsection draws heavily on Hellerstein and Neumark (2012).

“the social isolation of many residents in distressed areas” that “results in incomplete knowledge of the labor market and limited exposure to people in the labor market who may serve as the informal contacts needed for successful job searches” (p. 196). This is largely an unexplored area of research, although a study of Jobs-Plus provides some of the first evidence of which we are aware (Riccio, 1999). Jobs-Plus aimed to increase labor supply incentives for public housing residents by reducing the rent increases that accompany increases in earnings. In addition to including employment-related activities and services, Jobs-Plus endeavored to encourage the formation of labor-market networks or to provide functions similar to those supplied by networks. Most sites had “job developers” on staff whose responsibilities included providing outreach to local employers, cultivating relationships with them in an effort to place Jobs-Plus participants in employment (Kato et al., 2003). The program also employed residents as “court captains” or “building captains” who maintained contact with other participants, including sharing information about employment opportunities.

Jobs-Plus had a clear place-based flavor, attempting to transform the community through a saturation strategy that targeted all nondisabled working-age residents of these projects, rather than just trying to change individual behavior. This was based on the network-related (and peer effect-related) theory that saturation can lead to tipping points, creating a critical mass of employed residents who succeed in the workforce. These employed residents would, it was hypothesized, “signal to others the feasibility and benefits of working, elevate and strengthen social norms that encourage work, foster the growth of work-supporting social networks, and . . . contribute to still more residents getting and keeping jobs” (Riccio, 1999, p. 13).

There is evidence that the program delivered economic benefits in terms of both earnings and employment (Bloom et al., 2005). It is difficult, however, to draw firm conclusions about the value-added of specific efforts to build labor-market network connections due to two problems: first, the implementation of the network component of Jobs-Plus was spotty and encountered unanticipated difficulties, and second, it is hard to tell which components of the Jobs-Plus program delivered economic gains to its participants.

The reports on Jobs-Plus are replete with discussions of the problems encountered regarding building and strengthening networks. For example, Kato et al. (2003) noted that community support for work was the slowest component to develop (p. iii) and one site never developed it (p. 3). Bloom et al. (2005) noted, with regard to community support for work, that “[a]lthough many of these kinds of activities were tried at some points during the demonstration . . . most did not take root. What did take root—and grow—was the idea of using a small group of residents as extension agents of Jobs-Plus” (p. 48). Issues also arose regarding difficulties in doing outreach owing to high levels of illicit activity in some developments. Residents interested in working sometimes reported that a desire to stay out of trouble, combined with criminal activity among residents, “discouraged them from interacting with other residents, for fear that their

neighbors might be complicit” (Kato, 2004, p. 30). More to the point with regard to networks, there was a concern that someone you might refer could reflect badly on you.⁴⁷ Despite these difficulties, however, the description of implementation reveals numerous cases of job developers and sometimes captains developing means of linking residents to employment opportunities, likely providing labor-market contacts that many of the participants were lacking.

It is, however, difficult to attribute the gains from the Jobs-Plus program specifically to the network efforts, because the evaluations focused on the overall success of the program, rather than parsing out the effects of the separate components. Thus, at the end of the day, it is difficult to point to the evidence from Jobs-Plus as establishing the productivity of place-based policy efforts focusing on strengthening labor-market network ties. It is possible that most of the effect of Jobs-Plus came from the increased financial incentives to work generated by restructuring the effects of earning increases on rent or other components of the program.

The discussion in this section thus far has focused on how network effects might be leveraged to *help* residents of low-income, urban labor markets. There is, however, an alternative perspective, whereby networks can diminish the effectiveness of other policies. For example, as noted earlier with respect to enterprise zones, such policies may be ineffective at improving local labor markets because businesses may not hire locals in these neighborhoods. Dickens (1999) echoed this concern (p. 394), citing some case-study evidence from Kasinitz and Rosenberg (1996) suggesting that employers may even *prefer* to hire those who live farther away, because employers were worried that locals would have trouble avoiding family problems at work and could be pressured by locals to help burglarize them.

Thus, networks can be a two-edged sword, not only reducing labor-market search frictions by increasing the flow of information about jobs but also potentially introducing rigidities by making continued hiring within a particular network lower cost than going outside the network, even though the latter may be necessary to deliver benefits to residents of target areas. We believe that the design and evaluation of place-based policies that try to leverage labor-market networks to improve job market outcomes for local residents are a high priority for research and policy alike.

18.5.3 Discretionary grant-based policies

The specific aims of discretionary grant-based policies often include attracting new firms and investment to an area and either job creation or the prevention of job losses. Research

⁴⁷ This parallels findings from more systematic research by Smith (2005), who, based on in-depth interviews of low-income blacks, concludes: “Over 80 percent of respondents . . . expressed concern that job seekers in their networks were too unmotivated to accept assistance, required great expenditures of time and emotional energy, or acted too irresponsibly on the job, thereby jeopardizing contacts’ own reputations in the eyes of employers and negatively affecting their already-tenuous labor market prospects” (p. 3).

has addressed whether these policies achieve these aims by evaluating effects on the targeted outcomes for subsidized firms. In some cases, a somewhat broader analysis of the welfare effects of these programs has been attempted, for example looking at effects on the productivity of participant firms and of other firms whose performance might be affected by agglomeration externalities, as well as looking at employment displacement across areas. Much of this research faces a similar problem to that confronted by the literature on enterprise zones—in this case, that the eligible areas, in particular for countries in the European Union, are simultaneously eligible for other sources of regional development assistance. It is therefore crucial to control for these other sources of support to isolate the effects of discretionary grant policies.

A number of papers have examined whether these subsidies are effective in attracting new entrants. Two, for France and the United Kingdom, estimate location choice models using firm microdata and find some evidence of statistically significant but very small effects of discretionary grants on firm location decisions. [Crozet et al. \(2004\)](#) estimated conditional logit and nested logit models to examine the effect of the Prime d'Aménagement du Territoire (PAT) on foreign multinational location decisions across NUTS3 areas (départements) in France.⁴⁸ PAT grants aim to create or safeguard employment in lagging regions, and the authors reported that around half of annual expenditure under this scheme went to foreign-owned firms. The authors measured the generosity of the program at each potential location using data on the allocation of grant funding at the broader NUTS2 region-year level. While they find some evidence of positive effects of these grants on location choice, this is not highly robust, and even where there is a positive effect, its magnitude is dwarfed by that of other factors such as market access and agglomeration externalities. They also examine the effect of EU regional policy funds (covered in [Section 18.5.5](#)), which may affect the attractiveness of locations for investment, but find no evidence that this EU expenditure acted to boost the appeal of the qualifying locations for foreign investors.

[Devereux et al. \(2007\)](#) analyzed the effect of the similar UK RSA scheme on firm location decisions. The policy provides subsidies to firms for new investment in physical capital, with an ultimate aim of creating and safeguarding jobs and attracting FDI. The scheme aims to subsidize only additional investment that is attributable to the funding. Firms apply for a subsidy under the scheme and the application is assessed before a subsidy offer—conditional on job targets being met—is made. The funding is only available in designated Assisted Areas, and the rate of the subsidy is governed by EU rules and varies

⁴⁸ The European Commission defines three levels of regional units, under the nomenclature des unités territoriales statistiques (NUTS) classification. NUTS1 regions are the largest with populations of 3–7 million, for example, England, Scotland, and Wales, within Great Britain. NUTS2 regions are groups of administrative regions within countries containing 0.8–3 million inhabitants, while NUTS3 regions contain 150–800 thousand inhabitants.

with area characteristics. While Crozet et al. used area-level data on grant expenditure, Devereux et al. used microdata on the value of subsidies offered to individual firms to form a prediction of the subsidy that each new entrant might expect to receive in each location according to their characteristics.⁴⁹ They then use these predicted subsidy offers in a conditional logit location choice model across counties within Great Britain. The study finds that the grants had a positive and statistically significant effect on the location choice of new entrants (both multinationals and new plants established by existing UK manufacturing firms). However, the magnitude of the effect was extremely small. Grants were found to be more effective in locations that had greater preexisting employment in the relevant industry—i.e., their impact on location choices was magnified in areas that might also benefit from localization economies—but the overall effect remained very small.

Overall, the evidence from both studies suggests that such incentives have little leverage in terms of influencing where firms choose to locate, a conclusion that has implications for the use and cost-effectiveness of this type of policy as a tool to try and create or enhance industrial clusters. In addition, it also suggests that unlike the evidence for enterprise zones, this type of discretionary, plant-specific and potentially heavily monitored grant policy is unlikely to result in significant relocation of activity into eligible areas. Indeed, in the United Kingdom, the policy rules have explicitly included as a requirement that the grant is only awarded if it does not involve the displacement of employment, for example, across plants in different locations within a firm.

Criscuolo et al. (2012) evaluated the same policy as Devereux et al. (2007), but analyzed its effect on employment, investment, and productivity, including displacement effects across areas. Their IV evaluation strategy was outlined in Section 18.4.5. Their findings suggest that RSA grants are effective in generating both additional employment and investment but that they do not increase productivity. Using plant-level panel data, they find a positive effect of receipt of a grant on employment. The magnitude of the estimated effect increases when they instrument participation in the RSA program with their exogenous policy instrument—the maximum subsidy rate mandated by the EU, which varies over time and by area and is zero in noneligible areas. The downward bias in the OLS estimates is consistent with the idea that there is selection both into applying to the scheme and in the successful award of a grant, by firms facing negative shocks to performance. Their estimates suggest that participation in the RSA scheme increases plant-level employment substantially, by 43%, where the mean and median plant sizes in their sample are 79 and 6 employees, respectively. They also investigate heterogeneity in the effects of the policy across plants owned by firms of different sizes. Their results

⁴⁹ They also estimate a first-stage selection equation to address the fact that in practice only certain firms apply and that the expected grant for each new entrant in each location can only be estimated from data on a set of firms that applied to the scheme and were successful in receiving an offer.

imply that positive effects are confined to plants that are part of smaller firms with fewer than 150 employees. While this is potentially in line with the idea that smaller firms are more likely to face financial constraints, which a subsidy to capital investment can alleviate, it might also be that larger firms are better able to capture rents under the scheme.

They then estimate the impact of the policy on investment and productivity at the firm level. They find positive effects on investment, which are again restricted to smaller firms. Moreover, the estimated impacts on investment are greater than the impacts on employment, in line with the design of the policy, which provides a direct subsidy to capital. There is no evidence that the policy led to increased TFP or average wages at the firm level. Therefore, since their descriptive statistics of prior characteristics show that it is less productive firms that typically receive the subsidies and since the scheme acts to increase employment in these firms, this implies a negative effect on aggregate productivity. Given the design of the policy—a subsidy to capital combined with employment targets to either create or safeguard jobs—the authors rightly argued that while the subsidy may lead to an increase in capital per worker, there is no *ex ante* reason for it to lead to increased TFP.

Criscuolo et al. then analyzed impacts at the area level and found that the policy acted to increase employment both within existing firms and via new firm entry, and reduced unemployment in these lagging areas. They also found little evidence of displacement effects on nonparticipant firms or for noneligible but neighboring areas. The only evidence they found that is suggestive of displacement is for plants that are part of larger firms, which is in line with larger, multiplant firms having more flexibility to move plants or employment from outside to inside an eligible area. Since some of the areas in which these grants are available also qualify for EU Structural Funds, the authors included an indicator variable for areas receiving these transfers, but found no evidence that this alternative funding acted to increase employment. Moreover, controlling for the presence of this other source of regional support did not affect the estimates of the impact of the RSA scheme.

Although they did not conduct a full cost–benefit analysis of the policy, the authors performed some additional calculations on the magnitudes of their estimated effects. Their results suggest that a 10% subsidy rate increases area employment by nearly 2.9%. At the average subsidy rate of 24%, this translates to an increase of around 111,000 jobs over the period 1986–2004 (from a basis of 1.6 million employees in manufacturing), at an estimated cost per job of nearly £4900 in 2010 prices. They reported that this is lower than the previous estimates of the cost per job under the policy and argued that this is because their IV strategy overcomes the downward bias affecting OLS results. Further calculations for the estimated number of individuals drawn out of unemployment are close to those for the number of additional jobs created, again suggesting that the policy is primarily acting to reduce unemployment rather than displacing employment from other firms or areas.

A similar subsidy program known as Law 488 exists in Italy. The policy aims to create new jobs via a subsidy to investment projects, for example, start-up physical capital investment or modernization or relocation projects. Applicant firms must specify the number of new jobs to be created alongside details of the investment project. The creation of these new jobs is binding on recipient firms in that if fewer individuals are employed, the entire value of the subsidy must be repaid. In addition, the number of jobs to be created per unit of investment is one of the criteria used to score and subsequently rank each application in terms of priority for funding. The other two main criteria are the share of project costs borne by the firm and how high the rate of subsidy requested is relative to the maximum rate allowable in the area under EU regulations (Bernini and Pellegrini, 2011, p. 254).

Bernini and Pellegrini (2011) evaluated the impact of this policy on the performance of subsidized firms using a difference-in-differences estimator with unsuccessful applicants as controls. They considered a wide set of outcome variables, including growth in output, value-added, employment, capital investment, and measures of profitability and productivity. Using data from the year before the application for the subsidy and for the project completion year, they found strong evidence that while the subsidies acted to increase output, employment, and investment growth, growth in output increased to a lesser extent than employment, and both labor productivity and TFP growth declined. Their estimates suggest that, over on average a 3.6-year period, total output growth was around 8–10% higher in subsidized firms, while employment growth was around 16–17% higher. Growth in physical capital was even higher at around 40%. Growth in labor productivity as measured by output per worker was found to be 7% lower, and TFP growth 8% lower, as a result of the subsidies. Similar to Criscuolo et al., they also found evidence that the effects on output and employment growth were greater for small firms, some evidence of lower labor productivity growth, but no evidence of lower TFP growth for this group.

While the evidence implies that the program generated positive effects on the targeted outcomes of capital investment and employment, the negative effects on productivity may be due to distortions brought about by the policy design. The subsidy may have induced firms to undertake investment projects that would not have been funded by external capital markets. If firms were not subject to external financing constraints in the absence of the policy and were optimally not choosing to undertake investment with an insufficiently high return, then any genuinely additional investment brought about by the subsidy may have been relatively unproductive. The combination of a subsidy to capital investment coupled with an explicit target for job creation (and a potential incentive to overstate the number of new jobs plus a subsequent requirement to meet this target) may create distortions to firms' optimal capital–labor ratios. In addition, the design of the scheme may even have created incentives to recruit relatively low-productivity, low-wage workers to meet the binding employment targets, potentially implying

a trade-off between the stated employment goals of the program and any productivity objectives (although perhaps with positive distributional effects). It may also be the case that nonrecipient firms had stronger incentives to increase their own efficiency, potentially due to increased competition from the subsidized firms.

Bronzini and de Blasio (2006) investigated whether Law 488 generated additional investment, focusing on the time profile of investment in subsidized firms relative to unsuccessful applicants. They found evidence of an increase in investment in subsidized firms 2 years after the subsidy was awarded. This is exactly the point at which firms undertaking a 2-year investment project under the scheme would receive the second half of their funds and at which firms undertaking a longer project would receive the second third of the subsidy. The payment of these installments is conditional on firms having undertaken their specified investment projects, and hence, this is the exact point at which a positive impact would be expected. However, looking over a longer time horizon, they found that recipient firms show a decrease in investment relative to the control group at 5 years, i.e., 2–3 years after the final subsidy payment is made.

The authors interpreted this as evidence that the policy may not in fact have generated additional investment, but merely brought forward investment that would have happened eventually. Potentially in line with this argument, they found that at the 5-year point, subsidized firms exhibit lower debt, which is suggestive of firms bringing forward investment and substituting public funds for private borrowing that would otherwise finance investment at a later date. One issue is how to square the findings with those of Bernini and Pellegrini (2011), who found a larger effect on capital investment. One explanation is that the latter used information on the exact completion date of the project and did not look at investment beyond that point; by conducting the analysis over a slightly shorter time period, they miss a fall in investment relative to the control group after the subsidy has been paid.

Some policies aim explicitly at attracting very substantial new investment and influencing where large, new entrants locate. The PAT and RSA schemes discussed above have an element of this, for example, in using multimillion dollar investment grants to attract internationally mobile FDI. Greenstone et al. (2010) evaluated the effects of subsidies to attract large plants to specific US counties. Rather than focusing on the effects on the recipient plant, they looked for evidence that the new plant generates positive agglomeration externalities or productivity spillovers, the existence of which would potentially provide a rationale for the use of such subsidies. As discussed in Section 18.4.5, the study uses information on counties that narrowly missed out on winning the plant as counterfactual locations. The authors found convincing evidence that the arrival of these large new plants led to higher TFP among incumbent plants, higher net entry of new plants, and an increase in county-level labor costs, with the latter implying that part of the productivity spillover accrues to workers in the form of higher wages, rather than simply leading to increased profits for incumbent firms. Spillover effects on

incumbent TFP are identified by estimating difference-in-differences specifications of plant-level production functions, which compare TFP growth in incumbent plants before and after the plant opening across paired winning and losing counties. Effects on wages are estimated using similar difference-in-differences specifications on individual-level wage data, controlling for worker characteristics.

While the main aim of the paper is to credibly identify agglomeration externalities in the context of a large-scale plant opening, the paper also provides evidence on the benefits of the subsidies. However, since the authors did not use information on the value of the subsidies given to these plants, the findings are not used in a full cost-benefit analysis. The authors found that 5 years after the opening of a new manufacturing plant, compared to plants in losing counties, incumbent plants in the winning location had on average around 12% higher TFP (equivalent to moving from the 10th to the 27th percentile of the observed TFP distribution across counties), evidence that strongly supports the existence of externalities and of substantial efficiency benefits stemming from the use of such subsidies. The authors translated this average effect into an increase of around \$430 million in incumbent plant output after 5 years. The average effect on wages is estimated to be an increase of around 2.7%. Using an equivalent econometric specification for productivity spillovers that implies an effect on incumbent plant TFP of 4.8%, a 23% share of labor in total costs implies that the estimated overall cost effect via higher wages is around 13% of the TFP increase. These estimated effects reflect the direct result of the opening of the new plant, as well as indirect effects through subsequent plant entry and any changes in the industrial mix of the area, together with changes in competition for inputs that will also influence the extent of spillovers to incumbent plants and input prices.

The average productivity spillover effect is perhaps not the most useful parameter for a policymaker. The authors did not find that *all* winning locations benefit; indeed, for some counties, the estimated spillover effects are negative. The paper investigates how the magnitude of externalities varies depending on a set of measures of economic proximity between the subsidized plant and incumbents. These measures are intended to capture the classic channels through which theory and other evidence suggest that spillovers might be stronger, namely, labor-market pooling or thick labor markets at the industry level, knowledge spillovers, and input-output linkages (the presence of thick markets for intermediate inputs that can benefit firms in vertically related industries). The results support the existence of heterogeneous effects across incumbents with different characteristics relative to the new plant, but not via all three mechanisms. Spillover benefits appear to be stronger for incumbents in industries that share labor with the industry of the new plant and also for those that have technological connections, but not for incumbent plants in industries that have upstream or downstream linkages.

A further point the authors made is that if increases in labor costs occur across all industries within a winning county, but the size of productivity spillovers are

industry-specific, then the effects on incumbent plants' profits can be highly asymmetrical. The existence of such asymmetrical effects could then explain the attractiveness of certain locations for specific sectors and the observed longevity of industrial clusters. An additional issue with regard to the scope of benefits that is not covered in the analysis, but is relevant to policymakers concerned with place, is whether heterogeneous effects occur not only according to economic proximity but also geographic proximity to the subsidized plant, in terms of the spatial scale on which spillover benefits accrue.

If, as the results suggest, the externalities generated by new plant openings are heterogeneous across locations, a policy where local areas compete to attract such plants could be welfare-enhancing at an aggregate and at a local level. In these circumstances, the payment of subsidies to firms, which mean that they internalize the externalities they generate in their location decisions, can increase efficiency. However, from a cost-benefit perspective, this argument also relies on local policy authorities being able to accurately assess the scale of future benefits arising from the establishment of new plants and being able to judge how much to bid. As the results in Greenstone et al. indicate, the estimated productivity spillovers arising from these new plant openings have not always been positive.

18.5.3.1 Summary of evidence on discretionary grants

The literature looking at the effects of discretionary programs has been creative in dealing with the additional identification problem posed by the selection of businesses into the schemes, as well as dealing with issues of location-based selection common to place-based policies that almost by definition target a highly nonrandom set of areas (see [Section 18.4.5](#)). Overall, the available evidence on the effects of discretionary subsidies on their targeted outcomes of investment and employment in recipient firms is positive; [Table 18.3](#) provides an overview of the findings. The fact that plants that receive subsidy offers have their applications pass through an initial scrutiny process, and that the targeted outcomes are often heavily monitored and that payment of the subsidy is contingent on the job and/or investment targets being met, may explain why these policies appear more successful in achieving their stated goals than, for example, enterprise zone programs. But close monitoring of these discretionary schemes will come at an administrative cost, and it would be instructive to have further evidence on how long the additional jobs created actually survived once the monitoring period expired, to form a better assessment of their longer run impact on employment.

The evidence also suggests that the design of some schemes might itself create distortions to firms' optimal capital-labor ratios and to productivity, with associated welfare effects. Indeed, to the extent that the schemes are designed to finance marginal investment projects that, absent any capital-market failures, would not be backed by private-sector finance, the subsidized investment may be relatively unproductive. Overall, the evidence does not provide any hint that these grants acted to improve productivity at

Table 18.3 Summary of evidence on discretionary grants

Study	Country	Program	Results
Crozet et al. (2004)	France	Prime d'Aménagement du Territoire	Small, nonrobust effects of PAT subsidies on foreign multinational firm location decisions
Devereux et al. (2007)	United Kingdom	Regional Selective Assistance	Small effects on location decisions of foreign multinational firms and domestic multiplant firms Heterogeneity in the effectiveness of grants in influencing location choice; grants having a greater effect in areas with higher existing employment in the firm's industry
Criscuolo et al. (2012)	United Kingdom	Regional Selective Assistance	Positive effects on plant employment (43% increase in employment for participant plants) and firm investment, but restricted to plants that are part of smaller firms (<150 firm employees); no evidence of effects on firm TFP or wages Positive effects on employment and number of plants at the area level (a 10% subsidy rate increases area employment by 2.9%) and negative effects on unemployment (a 10% subsidy rate reduces unemployment by 6.9%) No evidence, on average, of employment or plant displacement from noneligible to eligible areas, but some evidence of displacement for plants that are part of larger firms
Bernini and Pellegrini (2011)	Italy	Law 488	Output growth in subsidized firms around 8–10% higher over on average 3.6 years, employment growth 16–17% higher, and growth in physical capital around 40% higher; labor productivity growth and TFP growth 7% and 8% lower, respectively
Bronzini and de Blasio (2006)	Italy	Law 488	Effects on output and employment appear to be greater for small firms Increase in investment over the initial 2 years following receipt of the subsidy, but at 5 years, recipient firms show a decrease in investment relative to controls; program may act to bring forward investment that might otherwise have occurred at a later date, rather than subsidizing additional investment
Greenstone et al. (2010)	United States	Location subsidies for large plant entry	Substantial effects on incumbent plant productivity in successful locations; incumbent plant TFP 12% higher after 5 years Heterogeneity in magnitude of TFP effects across industries and across locations Positive effect on county-level wages (2.7%)

recipient plants, and if, as in the case of the UK evidence, recipient firms had relatively low productivity to start with and then subsequently expanded, this implies a negative effect on productivity in the aggregate.

Evidence from the United States does, however, imply that the use of subsidies to influence the location of very large new plants can result in productivity spillovers to incumbent firms. It is likely that the areas offering subsidies in the United States are quite different to those typically targeted in a European context. Greenstone et al. reported that US counties that win a new plant have superior economic characteristics, for example, in terms of incomes, population growth, and labor force participation rates, compared with the rest of the country, which could reflect Glaeser and Gottlieb's (2008) suggestion, discussed in Section 18.2.7, that efficient place-based policies may increase disparities between areas. In contrast, within the EU, the opposite is true, with subsidies applying to more economically distressed areas. In fact, even the Greenstone et al. results illustrate significant heterogeneity in the magnitude of the estimated spillover effects both across locations and across industries within those locations, implying that any benefits of the policy that are external to the recipient firm and that might provide a theoretical justification for the use of such subsidies may be highly contingent on local economic characteristics.

18.5.4 Clusters and universities

Some countries have tried to use the placement or promotion of universities as a tool to enhance local economic development. It is easy to point out that world-leading universities are typically located in economically thriving regions and cities, and many of the most famous examples of high-tech clusters are explicitly linked to, or at least located in the vicinity of, prestigious research universities, such as Silicon Valley, and Silicon Fen around Cambridge University (see also Carlino et al., 2012). But causal estimates of the effects of university presence on local economic performance are relatively scarce, and what evidence there is—which is primarily based on established universities in relatively affluent regions—does not necessarily translate to a lagging-region context.

More generally, the observation that within a number of countries, employment in a wide range of industries is geographically localized, together with evidence that such clusters appear to generate productivity spillovers, has led to debate about whether there is a role for policy in promoting cluster development. Some countries, such as France, have explicitly adopted clusters policies, whereas others have emphasized industrial clustering as part of an enterprise zone program, for example, the 2011 program in England.

18.5.4.1 Clusters policies

This section discusses the scant evidence base on the effectiveness of cluster policy initiatives that aim to improve cluster performance, and on policies that, as a side effect, may have influenced cluster development. Martin et al. (2011) evaluated the effects of a policy

aimed at boosting the productivity of existing clusters in France. The program, Local Productive Systems (LPS), provided funds to aid collaboration and cooperation between groups of firms in the same industry in the same local area. Projects typically operated at the geographic level of a *département* (administrative areas) or a smaller employment area (local labor-market areas). An application to the program was normally made by a local public authority, and if successful, firms then signed up to participate. Projects involved the establishment of local initiatives, for example, to boost the profile of the cluster or to promote exports. The level of subsidy provided under the program was relatively small, at an average of around 40,000 euros per project. The policy ran from 1999 to the mid-2000s, when it was replaced in 2005 by a larger scale policy known as “competitiveness clusters.”

The authors used an administrative dataset to identify a set of firms that participated in the LPS scheme. They used difference-in-differences and matching approaches to estimate whether the policy had any discernible effects on firm performance, such as TFP and employment. They also provided evidence on the types of firms that chose to participate. The scheme was not explicitly aimed at supporting underperforming regions and industries, but in practice, this is what occurred. Although the LPS participant firms were relatively large, they were less productive, with the latter driven by industry and location characteristics. For example, the authors found that LPS firms tended to be located in areas eligible for the PAT grant program outlined in [Section 18.5.3](#) above.

The estimates provide little evidence of a positive effect of the policy on TFP in LPS firms and, if anything, point towards TFP falling postparticipation relative to the controls. In a separate examination of preprogram trends, the authors found that TFP in participating firms appeared to be declining relative to nonparticipants even before receipt of the funding, which suggests that this might have driven their decision to engage with the scheme. At best, the results suggest that participation might have brought about a temporary halt to this decline in TFP. The authors looked at other adjustment margins but found no effect on employment growth or on firm-level exports.

The authors acknowledged that there may be spillover effects from participant to nonparticipant firms at least within region, which may affect the firm-level estimates. They therefore carried out a second estimation exercise at the industry-*département* level. An industry-area is defined as participating in LPS if it contains at least one LPS firm. The results primarily confirm the picture from the firm-level estimates, the only difference being some evidence of an increase in exports, although this is not robust across specifications. The paper also finds no evidence that firm survival rates were higher in industry-areas receiving funding. Overall, the results paint a rather negative picture of the impact of the scheme. Potential explanations include that the program was too focused on lagging regions, the development of which was an objective of the government agency that administered the policy, possible rent-seeking activity by firms, and insufficient funding per project to generate discernible effects.

Falck et al. (2010) analyzed the effects of a cluster policy in the state of Bavaria in Germany, which was more directly focused on innovation in high-tech sectors. From 1999 to 2004, the Bavarian High-Tech Offensive targeted businesses operating in five technological sectors including life science, information and communication technology, and environmental technologies. Over this period, 1.35 billion euros were spent on initiatives aimed at improving firm performance and enhancing public-sector research in these fields, with around 50% of the budget spent on the latter. For example, funds were spent on public-sector research infrastructure with an explicit aim of enabling private-sector firms to access the facilities. In addition, expenditure was targeted at improving networks and cooperation between firms operating within the same technological field within the state, and state-backed venture capital funding was also made available. Other initiatives included science parks where innovative firms could locate rent-free.

The evaluation uses a difference-in-difference-in-differences design comparing changes in firm-level innovation outcomes before and after the program in affected industries in Bavaria to the before–after change in the same industries outside Bavaria, relative to the corresponding comparison of changes in nonaffected industries. The authors estimated effects on three outcomes: an indicator of whether or not a firm has introduced a product or process innovation, an indicator of whether an associated patent application was filed, and log R&D expenditures. The results imply a positive effect of the cluster initiative on the first two outcomes, but a negative effect on the third. The likelihood of innovating in the targeted sectors increased by around 4.6 percentage points from a baseline of roughly 50%, with the probability of filing for a patent increasing by 5.7 percentage points. However, the results imply that R&D expenditures fell substantially, by around 19%. Although at a first glance this appears a contradictory set of findings, Falck et al. suggested that because R&D expenditures are a measure of innovation inputs, the result could be interpreted as a reduction in the costs of innovation.

The paper also tries to address how the policy might have led to what appear to be improvements in firms' innovation productivity. Using answers to survey questions on the innovation process, they find that affected firms were less likely to report that their innovation activity was hampered by a lack of opportunities for cooperation with public research institutions, a lack of access to external expertise, or difficulties in finding R&D personnel. This suggests the program was successful in enabling private-sector access to public-sector scientific research and that it may have increased network connections between firms.

Some policies may indirectly affect cluster formation and development. Fallick et al. (2006) provided evidence that for the computer industry, aspects of California state law may have aided labor mobility, which in turn may have led to increased human capital externalities and knowledge spillovers as sources of agglomeration economies. Greater flexibility to move between employers should result in a more efficient allocation of human capital and talent across firms, and potentially as a by-product increase inter-firm diffusion of knowledge. This latter effect could reduce individual firms' returns to

investment in innovation and in employee human capital and provide an incentive for them to use noncompete contracts to try to reduce labor mobility and reduce knowledge spillovers to competitor firms. However, unique historical features of California state law make such contracts unenforceable.

In support of the influence of state law, the paper finds that month-to-month transition rates to a new job for college-educated men are higher in the Silicon Valley cluster and other computer clusters within the state, than they are in computer industry clusters outside the state of California. However, the same pattern is not found for other industries, for which there is no evidence of significantly higher job transition rates in California. The authors developed a model under which, if the nature of the innovation process is one in which success is highly uncertain but with high return, it may be optimal to have many companies working independently on the same problem, such as within a cluster, and then have an ex post reallocation of labor to the successful innovator. They argued that the characteristics of innovation in the computer industry fit with this model. While only offering circumstantial evidence, the results are in line with California state labor law acting to aid labor mobility and increase knowledge spillovers.

18.5.4.2 Universities

While [Carlino and Kerr \(2015\)](#) cover agglomeration and innovation more generally, here, we focus on evidence on the effects of universities on private-sector activity, including industrial clusters and wider effects on local employment and wages. The main empirical challenges in assessing the impact of universities on local economic conditions come from the fact that the location, characteristics, and scale of universities are nonrandom across geographic space; that factors such as business innovation and productivity and the industrial and skill composition of an area will very likely themselves influence university performance; and that unobservable local productivity shocks will likely affect both private-sector activity and university-sector activity.

[Kantor and Whalley \(2014\)](#) attempted to overcome these identification issues using an IV strategy to analyze the effect of knowledge spillovers from universities on noneducation-sector labor income in US urban counties. Their empirical analysis examines the relationship between noneducation-sector wages and university expenditure. As an instrument for the latter, they exploited exogenous variation brought about by national stock market shocks, which affect the market value of universities' endowments and, through this, university expenditure.

The authors used a county–industry panel from 1981 to 1996, with a basic regression specification relating the long difference, $t - x$ to t , in log non-education labor income at the county–industry level, to the long difference in per capita university expenditure at the county level, and a set of year fixed effects. The instrument for changes in per capita university expenditure is constructed as an interaction between the lagged market value of universities' endowments within a county (dated $t - x - 1$) and the change in the S&P

500 Index over the period $t-x$ to t . Since universities follow fixed expenditure rules, spending a set fraction of around 4–5% of the market value of their endowments per annum, shocks to the stock market will directly, and differentially, affect expenditure for each university, dependent on the initial value of their endowment. The national stock market shocks, and lagged university endowment levels, must also be exogenous with respect to growth in local noneducation labor income.⁵⁰

The IV estimates imply that an increase in university expenditure of 10% increases wages for workers in the noneducation sector by 0.8%. However, this relatively small average effect masks some heterogeneity. The authors distinguished responses across different types of industries, focusing on those with closer links to the university sector and across counties with more or less research-intensive universities. They found that industries that are technologically closer to university research (measured by the propensity for industry patents to cite university patents), and hence more likely to benefit from knowledge spillovers from higher education, receive higher spillovers, as do industries that employ a higher fraction of graduates, and industries with higher worker transitions into and out of the higher education sector. In addition, locations with more research-intensive universities (measured by the fraction of university students who are graduate level) generate higher spillovers. Their results suggest that the skill and industry mix of an area and the type of university it hosts are important determinants of the degree to which locations can both benefit from, and generate, externalities from higher education. This plausible pattern of heterogeneity in effects suggests that the results are more likely to be causal.

More descriptive evidence consistent with effects of universities on industrial clusters includes [Abramovsky et al. \(2007\)](#) and [Abramovsky and Simpson \(2011\)](#), who analyzed whether the presence of universities, and in particular highly rated university research departments within universities, is associated with the location of firms' R&D facilities. They exploited data from the UK Research Assessment Exercise, which is used to allocate the main publicly funded grant for research and which provides ratings of individual university research departments, with the highest rated departments deemed to be carrying out internationally leading, frontier research. The authors linked this information at the area level to data on private-sector R&D labs. They found that for some industries, but by no means all, the geographic distribution of R&D labs is skewed towards locations with highly rated, industrially relevant university research departments. For example, [Abramovsky and Simpson \(2011\)](#) found that pharmaceutical firms tend to locate their R&D labs disproportionately within 10 km of world-class chemistry research departments. Similar evidence is found for R&D service firms—many of which serve the

⁵⁰ The authors carry out robustness checks to address the issue of unobservables that might be correlated with past endowment values and affect future income growth: e.g., the possibility that firms are differentially affected by stock market shocks and that the location of particular types of firms is correlated with initial endowment values.

pharmaceutical industry. However, their results cannot be interpreted as causal. They found that many of these positive relationships can be explained by the presence of science parks. While it is quite likely that the science parks result from the universities' research standing and also from private-sector demand for premises in proximity to these universities, it remains difficult to untangle the direction of causality.⁵¹

One study that assesses the effects of a deliberate policy of using the geographic location of higher education as a lever for regional development is [Andersson et al. \(2004\)](#). From 1987, Sweden engaged in a substantial expansion of higher education that involved an increase in student enrollment, the establishment of new colleges, and four existing colleges gaining university status. Part of the aim was to make higher education less centralized and more available to students across all locations in Sweden including more remote regions, with a view to increasing participation. However, the policy could also have had effects on the demand for skilled labor in these regions through increased employment in the higher education sector and, depending on migration patterns, the supply of graduates in these locations. It may also have generated effects through innovation or human capital externalities from higher education to the private sector.

Andersson et al. estimated the effects of higher education presence on local labor productivity (output per worker) using a municipality-level panel. Identification comes from within-municipality variation over time in two measures of the scale of higher education in the area—the number of researchers employed at higher education institutions and the number of students enrolled. The authors differentiated between newer and older established institutions (the original six universities in operation) and considered whether effects vary with geographic distance. Overall, they found evidence in line with positive effects on local labor productivity from the expansion of higher education, that the effects associated with university researchers are greater than those associated with expansion of student numbers, that the effects are greater with respect to newer institutions for both measures of university presence, and finally that the effects are strongly spatially concentrated. The latter results imply that over half of the estimated productivity gains accrue within 20 km of the border of the institution's host municipality. However, the study is unable to disentangle the underlying causes of effects on labor productivity. In principle, one would want, for example, to be able to distinguish between effects driven by changes in the composition of the workforce (for example, through increased skill levels) or through externalities from spillovers from university research leading to higher noneducation-sector TFP.

⁵¹ Other evidence in a similar vein includes [Woodward et al. \(2006\)](#), who found a positive but fairly weak relationship between proximity to university research, measured by total university R&D expenditure in science and engineering, and numbers of high-tech start-ups. Unlike [Abramovsky et al. \(2007\)](#) and [Abramovsky and Simpson \(2011\)](#), who found considerable heterogeneity across industries, the Woodward et al. findings are consistent with knowledge spillovers from university research across a number of high-tech sectors.

The same authors go some way to addressing this in [Andersson et al. \(2009\)](#) and revisit the effects of the policy. They extend the analysis to look at innovation as measured by patents granted and look at aggregate effects on labor productivity and output. Given that the locations chosen for university expansion may have been related to underlying economic variables such as their scope for future productivity growth, the paper also presents IV estimates, instrumenting for university presence and scale with measures of preexisting facilities in the area including nursing schools and military facilities (since the buildings were used as sites for the new institutions), the fraction of the local population turning 18, and the fraction of voters voting for different political parties (some of which were strong supporters of university decentralization).

Their findings for labor productivity are in line with the previous paper, with gains if anything estimated panel data to be even more localized. In addition, they estimated panel data count models relating numbers of patents granted in an area, which might better capture knowledge spillovers from university research, to numbers of research staff employed at old and new institutions. The results suggest a positive relationship between the measures of investment in higher education and innovative outcomes. Effects for both labor productivity and innovation are found to be increasing in the fraction of individuals in the area who hold doctorates (measured contemporaneously). This suggests that the benefits of the decentralization policy were asymmetrical across areas according to their human capital endowments, although the latter may also have evolved endogenously because of the policy.

As a final exercise, the paper estimates the net innovation and productivity gains at the national level by constructing a no-decentralization counterfactual whereby researchers based at new institutions created post-1987 are allocated proportionately to preexisting institutions and recalculating the levels of innovation and productivity in each region. The results of this exercise imply an aggregate zero effect on patents generated, suggesting that the spatial redistribution of university research staff did not lead to any aggregate gain in innovative activity. However, using the labor productivity estimates, the authors backed out an estimate of GDP gains from the policy of between 0.01% and 0.10%. While the authors did not attempt to reconcile these findings, the findings would seem to imply that the aggregate GDP gains are driven by human capital improvements and agglomeration externalities working through channels other than purely innovation. However, the noncitation-weighted patent measure will likely be an imperfect measure of innovative activity, since different industries will exhibit highly different propensities to patent and not all patents are of equal value; hence, it remains possible that the policy did result in additional innovation in the aggregate.

18.5.4.3 Summary of evidence on clusters and universities

The evidence on higher education institutions suggests that areas do benefit from productivity spillovers but that these may be highly localized and also industry-specific and, in particular, arise in industries with closer technological links to university research

and in industries that employ a higher proportion of university graduates. Knowledge spillovers might benefit incumbent firms, but the evidence is also at least suggestive of university research facilities acting to attract high-tech innovative firms to these areas and hence acting as a basis for cluster formation with potential long-term benefits accruing—at least to certain industries—through these agglomerative forces. Many of these studies are based, however, on long-established universities, very likely in relatively affluent regions. But from the point of view of using universities as a policy tool for economic development, the evidence from Sweden points towards beneficial effects on local labor productivity—effects that do not appear to net out on aggregate across regions and that are potentially driven by increases in local human capital endowments. Hence, the evidence implies that investment in higher education and research, which have public-good elements to them, could generate long-term local effects. But questions remain about the optimal location of higher education investment, and to inform this, more evidence is needed, in particular on the precise channels through which higher education institutions affect local economic activity and how effects vary with local characteristics.

To some degree, high-tech firms do appear to internalize externalities from public-sector research in their location decisions, although science parks and other incentives may also be influential. The results on the impact of the Bavarian cluster initiative also suggest that, aside from seeking to influence firm location decisions, government intervention can also potentially overcome coordination failures and increase the returns to innovative activity, by seeking to bolster interaction between private-sector firms and improve private-sector access to public research facilities. However, this appears to have been both a highly targeted program, with funds flowing to both private firms and public-sector research, and an expensive program, certainly in comparison to the LPS policy adopted in France, which did not result in any evidence of beneficial effects on firm performance.

18.5.5 Infrastructure investment and other regional policies

The European Union has long embraced policies aimed at reducing disparities between regions across all member states, with funds being distributed to lagging regions even within nations with on average relatively high per capita income. The primary policy instrument is EU Structural Funds, comprising the European Regional Development Fund (ERDF) and European Social Fund (ESF), which aim to increase economic growth and create jobs in eligible areas. ERDF expenditure is typically on infrastructure, for example investment in energy, telecommunications, or linked to R&D, but can also include subsidies for investment by firms. Expenditure under the ESF is on initiatives to boost employment, such as training programs or projects to increase labor-market attachment. A third pot of funding, the Cohesion Fund, is available to whole countries with gross national income per capita less than 90% of the EU average. This program

funds investment in cross-national transport infrastructure (see Redding and Turner, 2015, for a discussion of the evidence on transport infrastructure and growth) and investment with environmental benefits. During the period 2007–2013, expenditure under the three programs accounted for around 35% of the total European Community budget (Becker et al., 2012), with ERDF expenditure of 201 billion euros accounting for over half of this.⁵² The largest amount of funding goes to Objective 1 regions with GDP per capita below 75% of the EU average. Within these lagging areas, European national governments are also permitted under EU regulations to offer discretionary subsidies of the type discussed in Section 18.5.3.

The United States also provides an example of a very large-scale regional development program, the Tennessee Valley Authority (TVA), the long-run impact of which is analyzed by Kline and Moretti (2014b). The authors not only evaluated effects on the targeted region but also estimated aggregate effects for the United States as a whole and examined outcomes after the funding is withdrawn. We first discuss this and related US policy and then turn to EU policy.

The TVA development and modernization policy involved substantial investment in public infrastructure including energy (electricity-generating dams), transport (road networks and canals), and new schools. The investment in electricity-generating capacity was part of a deliberate strategy to attract manufacturing activity to the TVA area. In geographic scope, it spanned four US states, covering nearly all of Tennessee and areas of Kentucky, Alabama, and Mississippi. The program began in 1933, with the highest expenditure occurring during the 1940s and 1950s. Kline and Moretti reported that federal expenditure totaled around \$20 billion (in 2000\$) from 1934 to 2000, with transfers per household during the early 1950s peaking at around 10% of average household income in the region.

Using preprogram data from 1930, the authors demonstrated that the set of counties covered by the program performed worse on a set of economic indicators compared to US counties in general and compared with other southern counties. The fraction of employment in agriculture was higher, and the fraction in manufacturing lower, manufacturing wages were lower, and literacy rates were also lower, indicating lower productivity and human capital. Given these systematic differences in the characteristics of the TVA counties, Kline and Moretti exploited the fact that regional authorities were proposed but due to political reasons never established in other parts of the country, to construct an additional set of control areas. To do this, they approximate the geographic boundaries of six potential regional authorities and verify that many of their pre-TVA economic characteristics, including prior trends in the share of employment in manufacturing, are much closer to those in the TVA counties.

⁵² See http://ec.europa.eu/regional_policy/thefunds/funding/index_en.cfm (viewed 7 January 2014).

The authors first looked at impacts on the TVA counties for a range of outcomes over different periods: 1940–2000, and two subperiods 1940–1960 and 1960–2000. The subperiod split is chosen because after 1960, federal transfers to the program were negligible. For the full time period, their most consistent findings are that the growth rate of manufacturing employment was higher (at around 5–6% per decade) and the growth rate of agricultural employment lower (around 5–7% lower per decade) relative to control areas. In addition, median family income growth increased around 2.5% per decade because of the policy. The results by subperiod indicate that the faster growth in manufacturing employment occurred during both periods, although the effect was around 3 times higher during the initial two decades (around 10–12% per decade during 1940–1960, as opposed to 3–3.5% from 1960 to 2000). In stark contrast, agricultural employment is found to have experienced substantially faster growth up to 1960 (around 11–12% per decade) but dramatically slower growth thereafter (around 13–17% lower per decade).

The results over the full period 1940–2000 paint a picture of public infrastructure investment in the TVA counties increasing the pace of industrialization, shifting employment out of agriculture towards manufacturing. That the authors found little evidence of wage increases in the manufacturing sector implies that labor supply was elastic, with new workers either moving to the area or switching from agricultural to manufacturing employment. The increase in family income growth therefore is driven by changes in the composition of employment, since wages in manufacturing exceeded those in agriculture. The authors attributed the differential response of manufacturing and agricultural employment growth post-1960 to the presence of agglomeration externalities in manufacturing, discussed next. These externalities continued to make the TVA counties an attractive location for new manufacturing activity even in the face of depreciation of the initial infrastructure investments following the withdrawal of federal funding, whereas the authors argued that faster growth in agricultural employment did not persist because agriculture may not exhibit the same external economies ([Hornbeck and Naidu, 2014](#)).

The paper then analyzes the aggregate effects of the TVA program—whether it generated benefits at the national level or whether the gains in the TVA counties came at the expense of other areas. The authors’ theoretical framework for analyzing national effects allows for two channels through which the policy can affect aggregate labor productivity: a direct effect whereby investment in public infrastructure acts to raise private-sector productivity and an indirect effect arising from the presence of agglomeration economies in manufacturing. As outlined in [Section 18.2](#), this latter effect cannot have a positive impact on the aggregate unless different areas exhibit heterogeneity in how responsive local productivity is to a change in agglomeration, i.e., the local agglomeration elasticity. If this elasticity is constant across regions, then a spatial redistribution of workers will generate no aggregate benefits. Their empirical evidence is supportive of the latter case.

Rather than estimate effects on labor productivity, the authors assumed perfect labor mobility and estimate effects on manufacturing employment, which will increase in the face of labor productivity-enhancing investment.⁵³ The paper estimates both the direct effect of the TVA investment on manufacturing employment and the indirect effect on manufacturing employment of agglomeration, as measured by the lagged density of local manufacturing employment. One issue is how to separately identify the direct and indirect effects, given that a direct effect that increases manufacturing productivity will increase manufacturing employment, which in turn will affect the density of manufacturing employment. The model is estimated using a county-level panel in first differences, with the dependent variable the change in log county-level manufacturing employment over a decade. The direct effect is identified from the inclusion of a dummy variable for TVA counties (not differenced). The indirect effect is identified from lagged changes in the density of manufacturing employment, which are instrumented using longer (two-decade) lags.

A second aim is to allow the agglomeration elasticity to vary flexibly across the distribution of the density of manufacturing employment. This is done using piecewise splines in manufacturing employment density, estimating spline functions for low, medium, and high sections of the distribution. The estimates using a spline in the log of manufacturing employment density measure the agglomeration elasticity with respect to manufacturing employment, and there appear to be no significant differences in the estimated elasticities across the three sections of the density distribution, with a 1% increase in density resulting in a 0.4–0.47% increase in manufacturing employment.⁵⁴

The estimated direct effect over the entire sample period 1960–2000 is positive but statistically insignificantly different from zero in the IV specifications, suggesting no evidence of differential manufacturing employment growth in TVA counties over this period. However, this conceals considerable variation in the estimated direct effect over three subperiods 1940–1960, 1960–1980, and 1980–2000. When the sample is extended further back in time, the results indicate that for the early years—1940–1960—the TVA policy resulted in a significant direct boost to manufacturing employment, but for the final two periods, the direct effect is estimated to be negative but insignificant. Hence, while the direct effect of the policy is felt in the period when substantial federal transfers were being made, the indirect effects on manufacturing employment and productivity

⁵³ Their data are decadal; hence, the assumption needs to hold at this frequency. In their model, the TVA investment increases firm productivity, which increases the wage, leading to an inflow of workers until in the longer run, the wage returns to its equilibrium level but at a higher level of manufacturing employment. This higher level of employment will only be permanent if either the productivity increase from the investment is permanent or the elasticity of productivity with respect to employment density is nonlinear.

⁵⁴ Their central estimate of the elasticity of productivity with respect to density is around 0.2, which is somewhat higher than the majority of elasticity estimates reported in [Melo et al. \(2009\)](#) across a range of studies, but not outside the overall range of estimates.

within the TVA counties, generated by agglomeration externalities, continued after the funding had been withdrawn, as evidenced by the continuing faster growth in manufacturing employment in the affected counties noted earlier.

As a final exercise, they use their estimates in a cost–benefit analysis from the point of view of the aggregate effect on the United States. Since the constant agglomeration elasticity implies that agglomeration benefits of the policy cancel out across US counties, the only benefit is the direct benefit to the TVA counties, by raising the productivity of the manufacturing labor force in that location during 1940–1960. They estimate the net present value of the benefits delivered by the program to be \$23.8 billion, which exceed the net present value of the federal transfers of \$17.3 billion.

In summary, for the TVA counties, once the federal subsidy program ended, the gains to agricultural employment that arose during the earlier period were eventually eroded. But increases in manufacturing employment brought about by investment in public infrastructure continued well beyond the policy end date, due to agglomeration externalities. The estimated direct effects of this infrastructure investment on aggregate manufacturing productivity are positive, with estimated benefits exceeding the program cost. However, based on the near-constant estimated elasticities of manufacturing employment with respect to manufacturing density, the estimates of indirect effects on manufacturing arising from agglomeration externalities are around zero in aggregate, with the positive agglomeration benefits that accrued to the TVA region offset by negative effects elsewhere.

Glaeser and Gottlieb (2008) discussed the effects of the Appalachian Regional Commission (ARC) that, beginning in 1963, disbursed considerable federal funding, primarily for transportation infrastructure but also for expenditure on health and education, to counties spanning a large geographic area from Mississippi to New York. They estimated effects on ARC counties using other nontreated counties within the same states as controls (excluding those within 56 miles of the coast). They found some evidence of a positive effect of the infrastructure expenditure on population growth over 1970 to 1980, but no statistically significant evidence of an effect over a longer period 1970 to 2000, and importantly, no statistically significant evidence of an effect on growth in per capita income. Their results are quite different to the findings of a previous study by Isserman and Rephann (1995), who primarily used comparisons of mean growth rates across matched treatment–control pairs (this time excluding as potential controls counties within 60 miles of ARC counties) and found large effects of the program on per capita income. While the stark differences in findings may be due to the alternative ways in which the authors addressed the problem of constructing a proxy for unobservable counterfactual outcomes, Glaeser and Gottlieb did acknowledge that the standard errors around some of their estimates are large enough that they cannot rule out substantial positive effects. In fact, their conclusion is that evaluation of this type of wide-ranging expenditure policy can prove very difficult, due to the funding being spread thinly across

a very large geographic area and over a long time period, and the difficulty of controlling for many other confounding factors that might affect economic growth over the longer term across such a wide region.

Becker et al. (2010, 2012) provided recent evaluations of the impact of major infrastructure investments under EU regional policy. They focused on the impact of Structural Funds on growth in employment and GDP per capita in Objective 1 regions. Becker et al. (2010) exploited data on NUTS2 and NUTS3 regions for three rounds of the Structural Funds program between 1989 and 2006. The authors used the fact that the NUTS2 area eligibility threshold is in principle a strict cutoff at GDP per capita below 75% of the EU average, in an RD evaluation strategy, arguing that those regions with GDP per capita close to the threshold will have ex ante similar characteristics, but only those below the threshold will qualify for funding as Objective 1 regions. Since in practice there are a few exceptions to defining eligibility at the NUTS2 level—a small number of NUTS3 regions received funding and a small fraction of NUTS2 regions have a treatment status that does not adhere to the strict eligibility rule—the paper implements a fuzzy RD approach and instruments regions' treatment status using the eligibility rule, and also conducts a robustness check whereby treatment status is defined at the NUTS3 level.

Their results imply a robust positive effect of Structural Funds expenditure on growth in GDP per capita, with a preferred estimate of around 1.6 percentage points per annum within a funding period.⁵⁵ However, in the vast majority of specifications, they found no effect on employment growth. The authors conducted a back-of-the-envelope cost-benefit analysis, which suggests that the program of transfers is cost-effective, generating a return of 20%, or a multiplier of 1.2. However, the confidence intervals around their estimates mean that they cannot reject a multiplier of only 1. They speculated that the findings of effects on per capita GDP, combined with no effects on employment, are driven by productivity gains from infrastructure investment but that any new job creation may only occur with a longer time lag. Overall, though, their results imply that the policy is cost-effective in increasing income in targeted regions.

Becker et al. (2012) extended this analysis to look at the relationship between the generosity of financial assistance under the program, i.e., the treatment intensity, and growth in per capita income. Their approach allows them to analyze whether the Structural Funds budget could have been redistributed across eligible EU regions to achieve higher growth in the aggregate and faster convergence. The underlying idea is that if the investment funded by the transfer payments exhibits decreasing returns, it is possible that

⁵⁵ Their evidence of a positive effect on per capita GDP growth in Objective 1 regions is also supported by Mohl and Hagen (2010). Table 1 in that paper provides an excellent summary of the findings from evaluations of the effects of EU Structural Funds, together with information on the data and econometric approaches. The majority of studies suggest positive effects on regional growth or on regional convergence, with a few exceptions.

funding beyond a certain level becomes inefficient. It is also possible that very low levels of transfers are simply ineffective in stimulating growth, with a minimum funding generosity required to generate a big push.

The authors used data on the intensity of transfers under the Structural Funds and Cohesion budgets over two EU funding rounds (1994–1999 and 2000–2006) at the NUTS3 level. They identified substantial variation in the intensity of transfers across regions, as measured by annual transfers as a fraction of a region's GDP in the year before the funding began. This ranged from 0.00009% of GDP for a region within Sweden to 29% of GDP for a region in Greece, with an average intensity of 0.756%. The paper estimates the effects of different degrees of treatment intensity using generalized propensity score estimation. This nonparametric method is an extension to continuous treatments of the propensity score matching approach to analyzing the impact of a binary treatment. The method allows the authors to estimate whether the treatment effect varies at different funding intensities, conditional on observable determinants of the treatment intensity itself (Hirano and Imbens, 2004).

The results confirm the findings of Becker et al. (2010) that on average the program generated positive effects on regional growth, but they also imply a nonlinear relationship between treatment intensity and growth in per capita income. From this, the authors can back out various thresholds: first, the “maximum desirable treatment intensity,” defined as the intensity beyond which they cannot reject a null hypothesis of zero effects on growth, and the second, the “optimal transfer intensity,” the level at which one additional euro of funding generates exactly one euro of additional GDP in the average region.

They found that around 18% of NUTS3 regions received funding beyond the estimated maximum desirable treatment intensity threshold of approximately 1.3%. For these regions, growth would not have been substantially lessened by a reduction in the generosity of funding to this level. In addition, they estimated that a redistribution of funds away from these regions and towards regions with lower funding intensities would be more efficient and could have increased average regional growth in per capita income by around 1.12 and 0.76 percentage points in the first and second funding rounds, respectively. They also found that around 36% of regions received transfers beyond the optimal transfer intensity of 0.4% of regional GDP. Redistribution across regions based on this lower threshold could have raised aggregate GDP growth but would have come at the cost of working against the regional convergence objective, since the redistribution would have been towards relatively prosperous areas. This suggests a trade-off between maximizing the aggregate efficiency of the program and potentially taking advantage of greater agglomeration externalities in relatively well-off regions, and the specified redistributive aim of the scheme. The authors also found no evidence to suggest that a minimum level of transfers is necessary to generate increases in per capita income growth, estimating positive effects of even small transfers.

In summary, the evidence from the analysis of both the TVA investment program and the EU Structural Funds implies that infrastructure investment can be cost-effective in delivering productivity growth in targeted regions and can act as a redistributive tool across locations. Questions remain about how long-lasting these effects are, with the direct benefits of the TVA program appearing to erode over time, and also about the precise mechanisms underlying the effects on growth. For example, this type of funding typically covers a wide range of public infrastructure investments, and it would be instructive to know the relative benefits of each in terms of their effects and whether these investments are complementary to each other in terms of increasing local growth. There is also evidence that infrastructure investment can result in agglomeration benefits for the targeted areas, although this evidence also seems to suggest that these may come at the expense of efficiency gains in other regions.

18.5.6 Community development and locally led initiatives

Finally, we turn to a small number of place-based policies that do not fit so neatly into the previous categories. First, in the United States, there are a number of programs that focus on real-estate development, but sometimes also have other components. Moreover, these often have some discretionary flavor. The discussion of enterprise zones above already referred to redevelopment areas. A common tool in these areas is to allow tax increment financing (TIF) whereby increases in property taxes owed as a result of appreciation (presumably stemming from redevelopment) are used to finance the debt incurred to engage in the redevelopment. TIFs and redevelopment are somewhat controversial, and we do not discuss them in any detail in this chapter as the research on them focuses nearly exclusively on implications for real-estate prices (e.g., [Weber et al., 2007](#)). Low-income housing programs similarly have a place-based, discretionary flavor. Among studies examining their effects—again, mainly on housing markets—are [Sinai and Waldfogel \(2005\)](#) and [Eriksen and Rosenthal \(2010\)](#).

A recent study ([Freedman, 2012](#)) examines the federal New Markets Tax Credit (NMTC) program, which not only concentrates on real-estate development but also devotes resources to economic development mainly through subsidizing capital for businesses through loans or preferential interest rates. The study examines the same kinds of labor-market outcomes examined in the literature on enterprise zones. For the period of this study (2002–2009), the NMTC provided \$26 billion in tax credits to investors making capital investments mainly in businesses located in moderately low-income neighborhoods. Freedman reported that around 70% of the funds go to commercial real-estate development and most of the rest goes to business development—mainly loans to firms. NMTC funds are channeled through Community Development Entities (CDEs), often banks or financial institutions, which have to meet several criteria including serving or providing capital to low-income communities and people. The channeling through the CDEs is what gives the NMTC a discretionary flavor, especially given that

only a very small fraction received the right to allocate tax credits in the years Freedman studied, via a competitive process. The tax credits flow to investors that make equity investments in the CDEs.

The study has many parallels to [Freedman's \(2013\)](#) study of the Texas enterprise zone program discussed earlier, although the focus is national. He used the same data and a similar research design, exploiting a discontinuity in eligibility for NMTC funds based on the main criterion that makes tracts eligible—having median family income in the 2000 Census below 80% of the state's median for nonmetro areas and below the greater of the MSA or state median for metro areas. This is not the only rule determining eligibility, so Freedman used a fuzzy design that instruments for actual NMTC credits with whether or not a Census tract is eligible based on this rule. He carried out many of the same kinds of analyses to validate the RD design as in the enterprise zone paper. He did not, however, consider overlap with enterprise zones, which could be important.

The evidence suggests that there is a discontinuous increase in NMTC investment at the threshold for eligibility based on median family income in the tract—about \$1 million more in NMTC investment than similar tracts that do not qualify and about 0.05 additional businesses receiving investment. Given that these amounts seem fairly small, it may be more plausible to believe that the effects Freedman found—discussed next—flow more from the real-estate development side of the NMTC. Using Census data, the main statistically significant effect Freedman found is for reducing the poverty rate, with estimates centered on a reduction of about 0.8 percentage point. He characterized this as a limited and costly effect, so that despite the small investment effects, the evidence implies that it costs about \$23,500 to lift one person out of poverty. At the same time, Freedman also found some evidence consistent with compositional changes, with a few of the estimates indicating increases in household turnover of about 0.75 percentage point. Such displacement effects could imply even higher costs to reduce poverty. However, unlike some of the work on enterprise zones, Freedman did not find evidence of an effect on median housing values, with the estimates very close to zero, which is less consistent with a compositional change towards higher income, higher skilled people. Inferring a direct impact on poverty of residents is also challenging because there is no statistical evidence of employment effects from the LEHD data. While the point estimates hover around 1.5%, the standard errors are 3 times as large. Freedman generously concluded that there is a “modest positive effect on private-sector employment” (p. 1012), and while a positive but insignificant effect does not imply no effect, this still seems too strong a conclusion. And compositional shifts, in and of themselves, could also lead to higher employment. Between the potential compositional shifts, the difficulty of understanding how such small amounts could have much impact, and the small impacts that occurred even if we rule out compositional changes, it is hard to attribute much success to the NMTC program.

Some place-based community development policies involve considerable local autonomy in terms of designing interventions and spending public money. Two

examples within Europe are the Patti Territoriali program in Italy, which aimed to stimulate growth and employment, and the New Deal for Communities in England, which focused on community development with a very broad remit. Accetturo and de Blasio (2012) evaluated the effects of the Patti Territoriali program, established in 1997. These territorial pacts aim to boost economic growth and employment in lagging areas eligible for support under the EU criteria. During the period analyzed, in principle, the program covered the whole of southern Italy as an Objective 1 region and some areas of central and northern Italy. Within these areas, local governments, local business groups, and trade unions from proximate municipalities could come together to form an agreement, which set out a development plan for the area. Therefore, not all eligible municipalities participated in the program in practice, although the authors reported that those that did tended to form large groups covering on average 27 municipalities and an average population of 235,000. Public funding was allocated to both public infrastructure investment and financial incentives for private-sector investment in the participating areas. A maximum of 50 million euros was allocated to each area, with public infrastructure investment expenditure limited to 15 million euros.

The paper evaluates effects on employment and on plant numbers. Clearly, both the eligible municipalities and within those the subset that actually choose to participate are nonrandom samples of the full population. Within northern and central Italy, not all areas are eligible to form Patti Territoriali. The authors used propensity score matching to identify similar eligible and noneligible municipalities prior to the implementation of the program and used difference-in-differences to estimate an intention-to-treat effect, which can be a valuable parameter for policymakers wanting to know the effect of the program on the target areas. They also estimated the effect of treatment on the treated by comparing a set of participating areas with a set of comparable areas that were ineligible to participate.

These evaluation approaches are not available in assessing the impact of the program in southern Italy since the entire geographic area is eligible as an Objective 1 area. Instead, difference-in-differences estimates compare changes in outcomes in participating municipalities to changes in nonparticipating municipalities in the south. Equivalent difference-in-differences estimates, using eligible but nonparticipating municipalities as controls, are calculated for northern and central Italy to try to assess the degree to which selection into participation on unobservables might affect the estimated program effects for the south. The authors also addressed the fact that the Law 488 program discussed above was running concurrently.⁵⁶

⁵⁶ For northern and central Italy, only areas eligible for Patti Territoriali are eligible for Law 488 financial incentives, potentially biasing estimates of the effects upward. For southern Italy, the nonparticipant areas used as controls are also eligible for Law 488, which could bias estimates of the effects upward or downward depending on how Law 488 affected the different areas.

Using Italian census data from 1996 to 2001, the intention-to-treat estimates for central and northern Italy on employment and the number of plants are negative and not statistically significant. While it is possible that this reflects a combination of positive effects on participating areas and negative spillover or displacement effects on eligible but nonparticipating areas, the estimated effect of the program on participating municipalities is also negative and statistically insignificant. Estimates for the south, which compare outcomes for participating versus nonparticipating but eligible municipalities, imply a positive and significant effect of the program on employment. However, replicating this approach for the northern and central regions suggests that the estimates for the south are very likely upward-biased, and if the sample for the south is reduced to only those municipalities that received no funding under the Law 488 program, the estimated effect on employment decreases in size and becomes statistically insignificant.

These results point towards there being no positive effects of the policy on employment or on the creation of new plants. This raises the question of why it was not successful. One possibility the authors discussed is that the available funding of 50 million euros per area was spread too thinly to generate sufficient additional activity in these lagging regions. However, since this level of funding was found to be equally ineffective in the most deprived southern regions and in the relatively more prosperous regions of northern and central Italy, where perhaps a lower level of expenditure might have been required, the expenditure cap was likely not the only explanation. The second suggestion is that the program fell victim to rent-seeking activities, which could even have been heightened by the bottom-up, locally led approach, and that subsidies were diverted to inferior private-sector projects.

A second example of a policy with local autonomy is the New Deal for Communities, which was operational in England from 2002. The ultimate aim of the policy was to improve living standards in the most deprived neighborhoods in the country. In practice, the program involved local committees devising and implementing a range of policies that aimed to improve employment and educational attainment, reduce crime, improve health, and address local housing and environmental issues. Examples of projects aimed at increasing employment included advice and credit schemes for those wanting to start their own business, become self-employed, or develop an existing business, and support for vocational training. Thirty-nine neighborhoods participated in the scheme and a total budget of £2 billion was allocated to be spent supporting these local initiatives over a 10-year period.

[Gutiérrez Romero \(2009\)](#) used a difference-in-differences approach to analyze the effects of the program on employment outcomes in participating areas. Control group neighborhoods were selected based on being within the same—in principle eligible—local authority as treated neighborhoods, but not directly bordering the treated neighborhoods where the program was in operation. Using a household survey carried out in treatment and control neighborhoods, she found that the program increased the

likelihood of entering employment for specific types of individuals, such as those who were in full-time education or undertaking training in the preprogram period and those who were claiming incapacity benefits, but not for those who were claiming unemployment benefits. Partly in line with these results, in a companion paper using administrative data on benefit claimants, [Gutiérrez Romero and Noble \(2008\)](#) found that the same program led to a reduction in individuals in treated neighborhoods claiming unemployment and incapacity benefits.

Like in the analysis of enterprise zones, one point to come out of studies of local initiatives that include a very wide range of policy elements is the need to understand which components of these programs make a difference, in those instances where beneficial effects are detected. In addition, for policies that do not appear to generate effects, it would help to have further corroborative evidence on why they did not work—whether it is simply that the financial scale of the intervention is too low with funds being spread too thinly or whether the policy design was ineffective. In addition, the analyses of the NMTC and New Deal for Communities highlight the value of trying to isolate which sets of individuals are affected by the policy and whether the policy is reaching the target groups.

18.6. UNANSWERED QUESTIONS AND RESEARCH CHALLENGES

In summary, what have we learned from the available evidence? The answer is probably “not enough.” To guide policy, we need know more about *what* works, *why* it works, and, crucially for place-based policy, *where* it works and *for whom* it works. We conclude by suggesting five areas where the evidence base could be usefully extended: investigating long-run effects, isolating specific features of policies that make them effective or that create unwelcome distortions, identifying more precisely what the effects are and who it is that gains benefits or incurs costs, learning more about potential strategic interactions between jurisdictions offering place-based policies, and examining whether broader policy levers such as tax policy might be more effective than place-based initiatives.

In our view, a major shortcoming of the research on place-based policies is that even the most positive evidence on their effectiveness does not establish that they create self-sustaining economic gains. That is, at best, the evidence (sometimes) says that when place-based incentives *are in effect*, there are increases in economic activity and perhaps welfare. There may be some gains from benefits even if governments have to continue paying the costs. However, a much stronger case would exist if some kinds of place-based policies helped to jump-start economic development in an area in a way that becomes self-sustaining—in the language of economics, by moving the area to a new equilibrium. [Moretti \(2012\)](#) concurred, arguing “The real test is not whether [place-based policies] . . . create jobs during the push . . . Instead we need to look at whether the publicly financed seed can eventually generate a privately supported cluster that is large enough to become self-sustaining” (pp. 200–201).

Some of the most positive evidence summarized above seems to point towards the benefits of infrastructure expenditure, perhaps within enterprise zone-type programs, and more clearly as part of EU Structural Funds, as well as benefits from expenditure on higher education and university research. This is perhaps not surprising given the public-goods nature of this type of investment. However, more evidence is needed on longer term outcomes. The findings from [Kline and Moretti \(2014b\)](#) on the effects of investment under the TVA program suggest that this very substantial push did generate long-lasting increases in manufacturing employment and income in the targeted region, both through a direct effect during the peak period of infrastructure spending and through subsequent agglomeration externalities. But even this study finds that the direct effect of the public investment on local productivity diminished over time and also that when measured at the national level, the indirect agglomeration benefits to manufacturing appear to net out across regions at around zero. This analysis further highlights the value of assessing the aggregate welfare effects of these investment programs, and not just those on the directly affected areas, in order to fully determine the magnitude of any trade-off between aggregate efficiency and redistribution across regions—although that is very hard to do.

As summarized above, for enterprise zone policies, much of the evidence to date is at odds and presents mixed messages as to the effects of such programs. More could be done to try to reconcile the existing findings and to unpack whether differences are due to data, to econometric methods, or to genuine differences in the effectiveness of different programs operating in areas with different economic characteristics. Moreover, different features of programs may alter their effectiveness, as exemplified by the discussion of the potential role of block grants in the federal Empowerment Zones. This highlights our second point, that even if the current evidence did all point in the same direction—for example, in terms of positive employment effects—what it does not tell us is exactly which features of these programs matter—for example, the use of hiring credits versus infrastructure investment. Knowing more precisely what works in terms of specific elements of the policy, and importantly why these elements work, would be of considerable value to policymakers.

In addition, we need to know about how generalizable the policy conclusions are across areas. By their very nature, place-based policies are implemented in locations with different characteristics, and it is not only the policy details that vary but also the economic environments in which they are set. A few studies have looked for and found evidence of heterogeneous effects of the same policy across different locations—perhaps most notably and systematically the [Briant et al. \(2012\)](#) study of variation in the effectiveness of the French ZFUs based on transportation accessibility and barriers between targeted areas and main employment centers. Clearly, much more could be done to understand the source of this variation in policy effectiveness.

We have argued that studies that provide evidence on the outcomes specifically targeted by policies are valuable in their own right and are the obvious first step in policy

evaluation. However, research could look at a wider range of outcomes relevant to assessing overall welfare effects. For example, are there distortionary effects of enterprise zone policies beyond the displacement of employment from neighboring areas? If firms receive targeted incentives to increase employment, does this create any unwelcome distortions to firm productivity, for example, through the type of workers recruited or through the displacement of firms from otherwise more productive locations? While some of the empirical literature on discretionary grants assesses potential distortions of this type stemming from the policy design, the literature on enterprise zones does not typically look at plant labor productivity or TFP, although it has looked at related outcomes for individuals such as wages.

Research could also aim to shed more light on exactly who gains any benefits from place-based policies. If programs are effective in increasing local productivity, are the ultimate beneficiaries actually landowners if the supply of housing or buildings is inelastic? If policies are found to be effective in raising employment rates or average incomes in targeted areas, is it resident individuals in low-income groups, whom the policies often aim to benefit, who actually realize these gains, or is there significant in-migration? More evidence on the redistributive effects of these policies across individuals within eligible areas would be valuable, in particular linked to features of the programs such as hiring incentives for specific groups.

Some of the empirical literature pays attention to overlapping place-based policies, but that may be just the tip of the iceberg. Given the evidence that place-based policies often encourage simple relocation of economic activity, it seems natural to think that—at least for place-based policies at a local level—jurisdictions may respond to the policies offered nearby. This is a common theme in concerns about a “race to the bottom” in welfare programs, environmental regulation, and tax policy (e.g., [Brueckner, 1998](#)). This can have potentially important implications for empirical work: for example, the estimated short-term impact of a place-based policy implemented in a particular jurisdiction may capture the partial equilibrium effect, rather than the general equilibrium effect once other jurisdictions have responded. While this issue has been taken up with respect to other state or local policies, it seems it could be fruitfully considered in the context of the kinds of place-based policies this chapter considers and is also relevant to the issue of whether such policies are in reality a zero-sum game.

Finally, we noted earlier that there are policies intended to boost local labor-market activity—such as lowering business taxes—that are outside the definition of place-based policies, because they do not favorably treat one area *within* a government’s jurisdiction. In light of some of the theoretical arguments against place-based policies discussed in this chapter, it may well be that policies that encourage economic activity *without* distinguishing between regions within a jurisdiction are more efficient. [Bartik’s \(2012\)](#) review suggests that the evidence clearly indicates that lower local average business taxes are associated with higher local labor demand, although the range of elasticities is so large

(-0.1 to -0.6) as to make cost–benefit calculations meaningless. He also argued that evidence on Michigan’s MEGA program, which provided marginal subsidies to businesses that export from the local economy, shows that it is much more effective (6 times more impact per dollar of foregone revenue) than cutting the overall business tax rate (Bartik and Erickcek, 2010). Bartik also noted that we know a lot about policies to increase the quality of labor supply and how this can have positive externalities on others (Moretti, 2004), but not nearly as much about how a regional development policy to increase human capital results in more jobs and higher quality jobs in the local economy.

We think an important research question is whether place-based policies, per se, are more efficient, or less distortionary, than broader local economic development efforts. We are not aware of work that tries to weigh the alternative approaches, but given the fairly weak or highly uncertain evidence regarding the effectiveness of at least some place-based policies, it is likely that broader policies are more efficient. At the same time, it is plausible that the broader policies fail to achieve some of the distributional goals of place-based policies, although as we have noted, the evidence that place-based policies achieve their distributional goals is itself far from clear.

ACKNOWLEDGMENTS

We are grateful to Gilles Duranton, Vernon Henderson, Patrick Kline, Stephen Ross, Kurt Schmidheiny, and Will Strange for their very helpful comments.

REFERENCES

- Abramovsky, L., Simpson, H., 2011. Geographic proximity and firm–university innovation linkages: evidence from Great Britain. *J. Econ. Geogr.* 11, 949–977.
- Abramovsky, L., Harrison, R., Simpson, H., 2007. University research and the location of business R&D. *Econ. J.* 117 (519), C114–C141.
- Accetturo, A., de Blasio, G., 2012. Policies for local development: an evaluation of Italy’s “Patti Territoriali”. *Reg. Sci. Urban Econ.* 42, 15–26.
- Alder, S., Shao, L., Zilibotti, F., 2013. Economic reforms and industrial policy in a panel of Chinese cities. UBS Center Working paper No. 5. University of Zurich, Zurich, Switzerland.
- Ananat, E., Fu, S., Ross, S.L., 2013. Race-specific agglomeration economies: social distance and the black–white wage gap. NBER Working paper No. 18933. National Bureau of Economic Research, Cambridge, MA.
- Andersson, R., Quigley, J., Wilhelmsson, M., 2004. University decentralization as regional policy: the Swedish experiment. *J. Econ. Geogr.* 4, 371–388.
- Andersson, R., Quigley, J., Wilhelmsson, M., 2009. Urbanization, productivity and innovation: evidence from investment in higher education. *J. Urban Econ.* 66, 2–15.
- Andersson, F., Haltiwanger, J.C., Kutzbach, M.J., Pollakowski, H.O., Weinberg, D.H., 2014. Job displacement and the duration of joblessness: the role of spatial mismatch. NBER Working paper No. 20066. National Bureau of Economic Research, Cambridge, MA.
- Åslund, O., Östh, J., Zenou, Y., 2006. How important is access to jobs? Old question—improved answer. *J. Econ. Geogr.* 10, 389–422.

- Bartik, T., 2003. Local economic development policies. Upjohn Institute Working paper No. 09-91. Upjohn Institute, Kalamazoo, MI.
- Bartik, T., 2009. What should Michigan be doing to promote long-run economic development? Upjohn Institute Working paper No. 09-160. Upjohn Institute, Kalamazoo, MI.
- Bartik, T., 2012. The future of state and local economic development policy: what research is needed. *Growth Change* 43, 545–562.
- Bartik, T., Erickcek, G., 2010. The employment and fiscal effects of Michigan's MEGA tax credit program. Upjohn Institute Working paper No. 10-164. Upjohn Institute, Kalamazoo, MI.
- Baum-Snow, N., Ferreira, F., 2015. Causal inference in urban economics. In: Duranton, G., Henderson, J.V., Strange, W. (Eds.), *Handbook of Regional and Urban Economics*, vol. 5. Elsevier, Amsterdam.
- Bayer, P., Ross, S., Topa, G., 2008. Place of work and place of residence: informal hiring networks and labor market outcomes. *J. Polit. Econ.* 116, 1150–1196.
- Becker, S., Egger, P., von Ehrlich, M., 2010. Going NUTS: the effect of EU structural funds on regional performance. *J. Public Econ.* 94, 578–590.
- Becker, S., Egger, P., von Ehrlich, M., 2012. Too much of a good thing? On the growth effects of the EU's regional policy. *Eur. Econ. Rev.* 56, 648–668.
- Bernini, C., Pellegrini, G., 2011. How are growth and productivity in private firms affected by public subsidy? Evidence from a regional policy. *Reg. Sci. Urban Econ.* 41, 253–265.
- Billings, S., 2009. Do enterprise zones work? An analysis at the borders. *Public Financ. Rev.* 37, 68–93.
- Bitler, M., Gelbach, J., Hoynes, H., 2006. What mean impacts miss: distributional effects of welfare reform experiments. *Am. Econ. Rev.* 86, 988–1012.
- Bloom, H., Riccio, J., Verma, N., 2005. Promoting work in public housing: the effectiveness of Jobs-Plus. Final report. MDRC.
- Bondonio, D., Greenbaum, R., 2007. Do local tax incentives affect economic growth? What mean impacts miss in the analysis of enterprise zone policies. *Reg. Sci. Urban Econ.* 37, 121–136.
- Bound, J., Groen, J., Kédzi, G., Turner, S., 2004. Trade in university training: cross-state variation in the production and stock of college-educated labor. *J. Econ.* 121, 143–173.
- Briant, A., Lafourcade, M., Schmutz, B., 2012. Can tax breaks beat geography? Lessons from the French enterprise zone program. Paris School of Economics Working paper No. 2012-22. Paris, France.
- Bronzini, R., de Blasio, G., 2006. Evaluating the impact of investment incentives: the case of Italy's Law. *J. Urban Econ.* 60, 327–349.
- Brueckner, J., 1998. Testing for strategic interaction among local governments: the case of growth controls. *J. Urban Econ.* 44, 438–467.
- Brühlhart, M., Bucovetsky, S., Schmidheiny, K., 2015. Taxation and cities. In: Duranton, G., Henderson, J.V., Strange, W. (Eds.), *Handbook of Regional and Urban Economics*, vol. 5. Elsevier, Amsterdam.
- Burnes, D., 2012. An empirical analysis of the capitalization of enterprise zone tax incentives into commercial property values (Unpublished manuscript).
- Busso, M., Gregory, J., Kline, P., 2013. Assessing the incidence and efficiency of a prominent place based policy. *Am. Econ. Rev.* 103, 897–947.
- Carlino, G., Kerr, W., 2015. Agglomeration and innovation. In: Duranton, G., Henderson, J.V., Strange, W. (Eds.), *Handbook of Regional and Urban Economics*, vol. 5. Elsevier, Amsterdam.
- Carlino, G., Carr, J., Hunt, R., Smith, T., 2012. The agglomeration of R&D labs (Unpublished manuscript).
- Chang, A., 2013. Tax policy endogeneity: evidence from R&D tax credits (Unpublished manuscript).
- Combes, P.-P., Duranton, G., Gobillon, L., Roux, S., 2010. Estimating agglomeration economies with history, geology and worker effects. In: Glaeser, E. (Ed.), *Agglomeration Economics*. The University of Chicago Press, Chicago, IL.
- Combes, P.-P., Duranton, G., Gobillon, L., Puga, D., Roux, S., 2012. The productivity advantages of large cities: distinguishing agglomeration from firm selection. *Econometrica* 80 (6), 2543–2594.
- Crane, R., Manville, M., 2008. People or Place? Revisiting the Who Versus the Where of Urban Development. Lincoln Institute of Land Policy, Cambridge, MA. http://www.lincolninst.edu/pubs/1403_People-or-Place (Land Lines, July; viewed 7 January 2014).

- Criscuolo, C., Overman, H., Martin, R., Van Reenen, J., 2012. The causal effects of an industrial policy. NBER Working paper 17842. National Bureau of Economic Research, Cambridge, MA.
- Crozet, M., Mayer, T., Mucchielli, J.-M., 2004. How do firms agglomerate? A study of FDI in France. *Reg. Sci. Urban Econ.* 34, 27–54.
- Damm, A., 2014. Neighborhood quality and labor market outcomes: evidence from quasi-random neighborhood assignment of immigrants. *J. Urban Econ.* 79, 139–166.
- Devereux, M., Griffith, R., Simpson, H., 2007. Firm location decisions, regional grants and agglomeration externalities. *J. Public Econ.* 91, 413–435.
- Dickens, W., 1999. Rebuilding urban labor markets: what community development can accomplish. In: Fergusson, R., Dickens, W. (Eds.), *Urban Problems and Community Development*. Brookings Institution Press, Washington, DC.
- Dowall, D., 1996. An evaluation of California's enterprise zone programs. *Econ. Dev. Q.* 10, 352–368.
- Dujardin, C., Selod, H., Thomas, I., 2008. Residential segregation and unemployment: the case of Brussels. *Urban Stud.* 45, 89–113.
- Duncan, G., Zuberi, A., 2006. Mobility lessons from Gautreaux and moving to opportunity. *Northwestern J. Law Soc. Policy* 1, 110–126.
- Duranton, G., 2011. California Dreamin': the feeble case for cluster policies. *Rev. Econ. Anal.* 3, 3–45.
- Duranton, G., Puga, D., 2004. Micro-foundations of urban agglomeration economies. In: Henderson, J.V., Thisse, J.-F. (Eds.), *Handbook of Regional and Urban Economics*, vol. 4. Elsevier, Amsterdam.
- Elling, R.C., Sheldon, A.W., 1991. Comparative analyses of state enterprise zone programs. In: Green, R.E. (Ed.), *Enterprise Zones: New Directions in Economic Development*. Sage, Newbury Park, CA.
- Ellwood, D., 1986. The spatial mismatch hypothesis: are there jobs missing in the ghetto? In: Freeman, R., Holzer, H. (Eds.), *The Black Youth Employment Crisis*. University of Chicago Press, Chicago, IL.
- Elvery, J., 2009. The impact of enterprise zones on residential employment: an evaluation of the enterprise zone programs of California and Florida. *Econ. Dev. Q.* 23, 44–59.
- Erickson, R., Friedman, W., 1990. Enterprise zones: 2. A comparative analysis of zone performance and state government policies. *Environ. Plann. C: Gov. Policy* 8, 363–378.
- Eriksen, M., Rosenthal, S., 2010. Crowd out effects of place-based subsidized rental housing: new evidence from the LIHTC program. *J. Public Econ.* 94, 953–966.
- Falck, O., Heblich, S., Kipar, S., 2010. Industrial innovation: direct evidence from a cluster-oriented policy. *Reg. Sci. Urban Econ.* 40, 574–582.
- Fallick, B., Fleischman, C., Rebitzer, J., 2006. Job-hopping in Silicon Valley: some evidence concerning the microfoundations of a high-technology cluster. *Rev. Econ. Stat.* 88, 472–481.
- Fieldhouse, E., 1999. Ethnic minority unemployment and spatial mismatch: the case of London. *Urban Stud.* 36, 1569–1596.
- Freedman, M., 2012. Teaching new markets old tricks: the effects of subsidized investment on low-income neighborhoods. *J. Public Econ.* 96, 1000–1014.
- Freedman, M., 2013. Targeted business incentives and local labor markets. *J. Hum. Resour.* 48, 311–344.
- Fu, S., Ross, S.L., 2013. Wage premia in employment clusters: how important is worker heterogeneity? *J. Labor Econ.* 31, 271–304.
- Givord, P., Rathelot, R., Sillard, P., 2013. Place-based tax exemptions and displacement effects: an evaluation of the *Zones Franches Urbaines* program. *Reg. Sci. Urban Econ.* 43, 151–163.
- Glaeser, E., 2007. The economics approach to cities. NBER Working paper No. 13696. National Bureau of Economic Research, Cambridge, MA.
- Glaeser, E., Gottlieb, J., 2008. The economics of place-making policies. *Brookings Pap. Econ. Act.* 1, 155–239 (Spring).
- Gobillon, L., Selod, H., 2007. The effect of segregation and spatial mismatch on unemployment: evidence from France. CEPR Discussion Paper No. 6198. London, UK.
- Gobillon, L., Selod, H., Zenou, Y., 2007. The mechanisms of spatial mismatch. *Urban Stud.* 44, 2401–2427.
- Gobillon, L., Magnac, T., Selod, H., 2012. Do unemployed workers benefit from enterprise zones: the French experience. *J. Public Econ.* 96, 881–892.
- Gobillon, L., Rupert, P., Wasmer, E., 2013. Ethnic unemployment rates and frictional markets (Unpublished manuscript).

- Greenbaum, R., Engberg, J., 2004. The impact of state enterprise zones on urban manufacturing establishments. *J. Policy Anal. Manage.* 23, 315–339.
- Greenstone, M., Hornbeck, R., Moretti, E., 2010. Identifying agglomeration spillovers: evidence from winners and losers of large plant openings. *J. Polit. Econ.* 118, 536–598.
- Gutiérrez Romero, R., 2009. Estimating the impact of England's area-based intervention 'New Deal for Communities' on employment. *Reg. Sci. Urban Econ.* 39, 323–331.
- Gutiérrez Romero, R., Noble, M., 2008. Evaluating England's 'New Deal for Communities' programme using the difference-in-difference method. *J. Econ. Geogr.* 8, 759–778.
- Ham, J., Swenson, C., İmrohoroglu, A., Song, H., 2011. Government programs can improve local labor markets: evidence from state enterprise zones, federal Empowerment Zones and federal Enterprise Communities. *J. Public Econ.* 95, 779–797.
- Hanson, A., 2009. Local employment, poverty, and property value effects of geographically-targeted tax incentives: an instrumental variables approach. *Reg. Sci. Urban Econ.* 39, 721–731.
- Hanson, A., Rohlin, S., 2013. Do spatially targeted redevelopment programs spillover? *Reg. Sci. Urban Econ.* 43, 86–100.
- Hellerstein, J.K., Neumark, D., 2003. Ethnicity, language, and workplace segregation: evidence from a new matched employer-employee data set. *Ann. Econ. Stat.* 71–72, 19–78.
- Hellerstein, J.K., Neumark, D., 2012. Employment problems in black urban labor markets: problems and solutions. In: Jefferson, P. (Ed.), *The Oxford Handbook of the Economics of Poverty*. Oxford University Press, Oxford.
- Hellerstein, J.K., Neumark, D., McInerney, M., 2008. Spatial mismatch vs. racial mismatch? *J. Urban Econ.* 64, 467–479.
- Hellerstein, J.K., McInerney, M., Neumark, D., 2010. Spatial mismatch, immigrant networks, and Hispanic employment in the United States. *Ann. Econ. Stat.* 99 (100), 141–167.
- Hellerstein, J.K., McInerney, M., Neumark, D., 2011. Neighbors and co-workers: the importance of residential labor market networks. *J. Labor Econ.* 29, 659–695.
- Hellerstein, J.K., Kutzbach, M., Neumark, D., 2014. Do labor market networks have an important spatial dimension? *J. Urban Econ.* 79, 39–58.
- Hirano, K., Imbens, G.W., 2004. The propensity score with continuous treatments. In: Gelman, A., Meng, X.-L. (Eds.), *Applied Bayesian Modelling and Causal Inference from Incomplete-Data Perspectives*. Wiley, Chichester.
- Holzer, H., 1991. The spatial mismatch hypothesis: what has the evidence shown? *Urban Stud.* 28, 105–122.
- Hornbeck, R., Naidu, S., 2014. When the levee breaks: black migration and economic development in the American South. *Am. Econ. Rev.* 104, 963–990.
- Ihlanfeldt, K., Sjoquist, D., 1990. Job accessibility and racial differences in youth employment rates. *Am. Econ. Rev.* 80, 267–276.
- Ihlanfeldt, K., Sjoquist, D., 1998. The spatial mismatch hypothesis: a review of recent studies and their implications for welfare reform. *Hous. Policy Debate* 8, 849–892.
- Isserman, A., Rephann, T., 1995. The economic effects of the Appalachian Regional Commission: an empirical assessment of 26 years of regional development planning. *J. Am. Plan. Assoc.* 61, 345–364.
- Jacobs, J., 1961. *The Economics of Cities*. Vintage, New York, NY.
- Jencks, C., Mayer, S., 1990. Residential segregation, job proximity, and black job opportunities. In: Lynn, L., McGeary, M. (Eds.), *Inner-City Poverty in the United States*. National Academy Press, Washington, DC.
- Kantor, S., Whalley, A., 2014. Knowledge spillovers from research universities: evidence from endowment value shocks. *Rev. Econ. Stat.* 96, 171–188.
- Kasinitz, P., Rosenberg, J., 1996. Missing the connection: social isolation and employment on the Brooklyn waterfront. *Soc. Forces* 43, 180–196.
- Kato, L., 2004. *Mobilizing Resident Networks in Public Housing*. MDRC.
- Kato, L., et al., 2003. *Jobs-Plus Site-by-Site: Key Features of Mature Employment Programs in Seven Public Housing Communities*. MDRC.
- Kline, P., 2010. Place-based policies, heterogeneity and agglomeration. *Am. Econ. Rev. Pap. Proc.* 100, 383–387.

- Kline, P., Moretti, E., 2014a. People, places, and public policy: some simple welfare economics of local economic development programs. *Annu. Rev. Econ.* 6, 629–662.
- Kline, P., Moretti, E., 2014b. Local economic development, agglomeration economies, and the big push: 100 years of evidence from the Tennessee Valley Authority. *Q. J. Econ.* 129, 275–331.
- Kolko, J., Neumark, D., 2010. Do *some* enterprise zones create jobs? *J. Policy Anal. Manage.* 29, 5–38.
- Kolko, J., Neumark, D., Cuellar Mejia, M., 2013. What do business climate indexes teach us about state policy and economic growth? *J. Reg. Sci.* 53, 220–255.
- Ladd, H., 1994. Spatially targeted economic development strategies: do they work? *Cityscape: J. Pol. Devel. Res.* 1, 193–218.
- Lee, D., Lemieux, T., 2010. Regression discontinuity designs in economics. *J. Econ. Lit.* 48, 281–355.
- Ludwig, J., et al., 2013. What can we learn about neighbourhood effects from the Moving to Opportunity experiment? *Am. J. Sociol.* 114, 144–188.
- Lynch, D., Zax, J., 2011. Incidence and substitution in enterprise zone programs. *Public Financ. Rev.* 39, 226–255.
- Marshall, A., 1890. *Principles of Economics*. MacMillan, London.
- Martin, P., Mayer, T., Mayneris, F., 2011. Public support to clusters. A firm level study of French “Local Productive Systems” *Reg. Sci. Urban Econ.* 41, 108–123.
- Matouschek, N., Robert-Nicoud, F., 2005. The role of human capital investments in the location decision of firms. *Reg. Sci. Urban Econ.* 35, 570–583.
- Mayer, T., Mayneris, F., Py, L., 2012. The impact of urban enterprise zones on establishments’ location decisions: evidence from French ZFUs. CEPR Discussion Paper No. DP9074. London, UK.
- Melo, P., Graham, D., Noland, R., 2009. A meta-analysis of estimates of urban agglomeration economies. *Reg. Sci. Urban Econ.* 39, 332–342.
- Mohl, P., Hagen, T., 2010. Do EU structural funds promote economic growth? New evidence from various panel data approaches. *Reg. Sci. Urban Econ.* 40, 353–365.
- Montgomery, J.D., 1991. Social networks and labor-market outcomes: toward an economic analysis. *Am. Econ. Rev.* 81, 1408–1418.
- Moretti, E., 2004. Workers’ education, spillovers and productivity: evidence from plant-level production functions. *Am. Econ. Rev.* 94, 656–690.
- Moretti, E., 2010. Local labor markets. In: Card, D., Ashenfelter, O. (Eds.), *Handbook of Labor Economics*, vol. 4B. Elsevier, Amsterdam.
- Moretti, E., 2012. *The New Geography of Jobs*. Houghton Mifflin Harcourt, Boston, MA.
- Mulock, B., 2002. *Empowerment Zone/Enterprise Communities Program: Overview of Rounds I, II, and III*. Congressional Research Service, Washington D.C.
- Neumark, D., Grijalva, D., 2013. The employment effects of state hiring credits during and after the Great Recession. NBER Working paper No. 18928. National Bureau of Economic Research, Cambridge, MA.
- Neumark, D., Kolko, J., 2010. Do enterprise zones create jobs? Evidence from California’s enterprise zone program. *J. Urban Econ.* 68, 1–19.
- Neumark, D., Schweitzer, M., Wascher, W., 2005a. The effects of minimum wages on the distribution of family incomes. *J. Hum. Resour.* 40, 867–917.
- Neumark, D., Zhang, J., Wall, B., 2005b. Are businesses fleeing the state? Interstate business relocation and employment change in California. *Calif. Econ. Pol.* 1. http://www.ppic.org/content/pubs/cep/EP_1005DNEP.pdf (viewed 8 January 2014).
- O’Keefe, S., 2004. Job creation in California’s enterprise zones: a comparison using a propensity score matching model. *J. Urban Econ.* 55, 131–150.
- Papke, L., 1993. What do we know about enterprise zones? *Tax Policy Econ.* 7, 37–72.
- Papke, L., 1994. Tax policy and urban development: evidence from the Indiana enterprise zone program. *J. Public Econ.* 54, 37–49.
- Patacchini, E., Zenou, Y., 2012. Ethnic networks and employment outcomes. *Reg. Sci. Urban Econ.* 42, 938–949.
- Pelligrini, G., Muccigrosso, T., 2013. Do subsidised new firms survive longer? Evidence from a counterfactual approach (Unpublished manuscript).

- Peters, A., Fisher, P., 2002. State Enterprise Zone Programs: Have They Worked? W.E. Upjohn Institute for Employment Research, Kalamazoo, MI.
- Raphael, S., 1998. The spatial mismatch hypothesis and black youth joblessness: evidence from the San Francisco Bay Area. *J. Urban Econ.* 43, 79–111.
- Redding, S., Turner, M., 2015. Transportation. In: Duranton, G., Henderson, J.V., Strange, W. (Eds.), *Handbook of Regional and Urban Economics*, vol. 5. Elsevier, Amsterdam.
- Reynolds, C., Rohlin, S., 2013. The effects of location-based tax policies on the distribution of household income: evidence from the federal Empowerment Zone program (Unpublished manuscript).
- Reynolds, C., Rohlin, S., 2014. Do location-based tax incentives improve quality of life and quality of business environment? *J. Reg. Sci.* 54, 1–32.
- Riccio, J., 1999. Mobilizing Public Housing Communities for Work: Origins and Early Accomplishments of the Jobs-Plus Demonstration. MDRC, New York, NY.
- Rosenbaum, J., Zuberi, A., 2010. Comparing residential mobility programs: design elements, neighborhood placements, and outcomes in MTO and Gautreaux. *Hous. Policy Debate* 1, 27–41.
- Rosenthal, S., Strange, W., 2003. Geography, industrial organization, and agglomeration. *Rev. Econ. Stat.* 85, 377–393.
- Rosenthal, S., Strange, W., 2004. Evidence on the nature and sources of agglomeration economies. In: Henderson, J.V., Thisse, J.-F. (Eds.), *Handbook of Regional and Urban Economics*, vol. 4. Elsevier, Amsterdam.
- Sinai, T., Waldfogel, J., 2005. Do low-income housing subsidies increase the occupied housing stock? *J. Public Econ.* 89, 2137–2164.
- Smith, S.S., 2005. ‘Don’t put my name on it’: social capital activation and job finding assistance among the black urban poor. *Am. J. Soc.* 111, 1–57.
- Topa, G., Zenou, Y., 2015. Neighbourhood versus networks. In: Duranton, G., Henderson, J.V., Strange, W. (Eds.), *Handbook of Regional and Urban Economics*, vol. 5. Elsevier, Amsterdam.
- US Government Accountability Office, 2006. Empowerment Zone and Enterprise Community Program, Washington, DC.
- Wang, J., 2013. The economic impact of Special Economic Zones: evidence from Chinese municipalities. *J. Dev. Econ.* 101, 133–147.
- Weber, R., Bhatta, S.D., Merriman, D., 2007. Spillovers from tax increment financing districts: implications for housing price appreciation. *Reg. Sci. Urban Econ.* 37, 259–281.
- Weinberg, B., 2000. Black residential centralization and the spatial mismatch hypothesis. *J. Urban Econ.* 48, 110–134.
- Wilder, M., Rubin, B., 1996. Rhetoric versus reality: a review of studies on state enterprise zone programs. *J. Am. Plan. Assoc.* 62, 472–492.
- Wilson, W., 1987. *The Truly Disadvantaged*. University of Chicago Press, Chicago, IL.
- Wilson, D., 2009. Beggar thy neighbour? The in-state, out-of-state and aggregate effects of R&D tax credits. *Rev. Econ. Stat.* 91, 431–436.
- Woodward, D., Figueiredo, O., Guimaraes, P., 2006. Beyond the Silicon Valley: university R&D and high-technology location. *J. Urban Econ.* 60, 15–32.
- Zhang, J., 2007. Access to venture capital and the performance of venture-backed start-ups in Silicon Valley. *Econ. Dev. Q.* 21, 124–147.

This page intentionally left blank

CHAPTER 19

Regulation and Housing Supply

Joseph Gyourko^{*,†}, Raven Molloy[‡]

^{*}The Wharton School, University of Pennsylvania, Philadelphia, PA, USA

[†]NBER, Cambridge, MA, USA

[‡]Board of Governors, Federal Reserve System, Washington, DC, USA

Contents

19.1. Introduction	1290
19.2. Data: Old and New	1294
19.2.1 Indirect measurement	1295
19.2.2 Building codes	1296
19.2.3 Land use controls	1297
19.2.4 Other measures	1302
19.2.5 Summary of measurement issues and concluding comments	1303
19.3. Determinants of Regulation	1304
19.3.1 Joint determination of land values and zoning	1304
19.3.2 Homeowners, developers, and local politics in a single community	1305
19.3.3 Supply of buildable land and historical density patterns	1310
19.3.4 Regulation in a multicomunity setting: Sorting and strategic interactions	1312
19.4. Effects of Regulation	1316
19.4.1 Effects on the price and quantity of housing	1316
19.4.2 Effects on urban form and homeownership	1322
19.4.3 Effects beyond housing markets	1325
19.5. Welfare Implications of Regulation	1327
19.6. Conclusion	1330
Acknowledgments	1333
References	1333

Abstract

A wide array of local government regulations influences the amount, location, and shape of residential development. In this chapter, we review the literature on the causes and effects of this type of regulation. We begin with a discussion of how researchers measure regulation empirically, which highlights the variety of methods that are used to constrain development. Many theories have been developed to explain why regulation arises, including the role of homeowners in the local political process, the influence of historical density, and the fiscal and exclusionary motives for zoning. As for the effects of regulation, most studies have found substantial effects on the housing market. In particular, regulation appears to raise house prices, reduce construction, reduce the elasticity of housing supply, and alter urban form. Other research has found that regulation influences local labor markets and household sorting across communities. Finally, we discuss the welfare implications of regulation. Although some specific rules clearly mitigate negative externalities, the benefits of more general forms of regulation are

very difficult to quantify. On balance, a few recent studies suggest that the overall efficiency losses from binding constraints on residential development could be quite large.

Keywords

Regulation, Housing supply, Zoning, Land use

JEL Classification Code

R31

19.1. INTRODUCTION

This chapter discusses the causes and consequences of local regulations that restrict land use or otherwise limit the supply of housing. Researchers studying housing, urban economics, and local public finance have devoted much attention to this topic because regulation appears to be the single most important influence on the supply of homes. In contrast, the homebuilding sector appears to require a relatively low cost of entry, as various economic censuses consistently report that there are well over 100,000 companies in the single-family construction business. Moreover, residential construction is not an industry dominated by a few large entities because the vast majority of firms do less than \$10 million in business per year. Hence, economists typically abstract from industrial organization-type considerations when modeling the housing supply.

Labor and material costs also do not appear to act as a major constraint on residential development. Some markets are unionized, while others are not, generating differences in the level of the labor component of construction costs across locations. More generally, [Gyourko and Saiz \(2006\)](#) documented a large degree of heterogeneity in structure production costs across local markets, which is correlated with a number of supply shifters including the extent of construction worker unionization, the level of local wages, the local topography as reflected in the presence of high hills and mountains, and the local regulatory environment as measured by an index of Internet chatter on construction regulations. Nevertheless, their data are consistent with the conclusion that the supply of structures is competitive in the sense that the production of any particular part of a house is a constant returns-to-scale activity that can be conducted at virtually any volume.

The light line in [Figure 19.1](#) shows further evidence on materials and labor costs by plotting the evolution of real construction costs from 1980 to 2013. Specifically, it depicts the construction costs, including all materials and labor costs, involved in putting up the physical structure associated with a modest-quality single-family housing unit that meets all relevant building code requirements. While there is local market variation around this aggregate time series, construction costs have been essentially flat in real terms over the past 30 years. This trajectory is consistent with the idea that any inelasticity

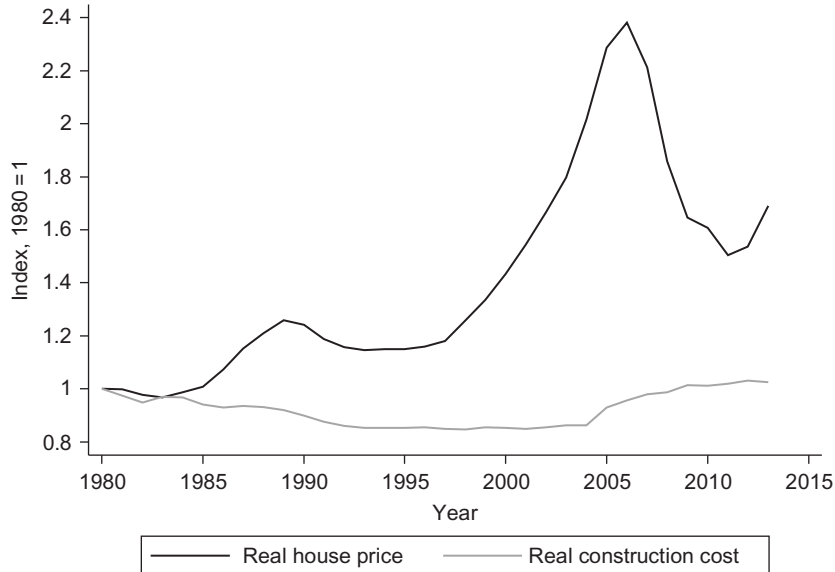


Figure 19.1 Real construction costs and house prices over time. *Note: Construction costs are the cost of an economy-quality home from RSMeans deflated by the consumer price index. House prices are the repeat-sales index published by CoreLogic deflated by the price index for personal consumption expenditures excluding housing services. The deflator was calculated by the authors from data published by the Bureau of Economic Analysis.*

of housing does not have at its root an inelasticity of the supply of the structure component of homes.¹

The land market is a very different story. In theory, the availability of buildable land might not constrain the supply of housing units if housing could be constructed as densely as necessary to meet demand. But in most places in the United States—and indeed around the world—local land use policy imposes limits on residential development that restrict the size and type of housing units that can be built on a given amount of land. These restrictions add extra costs to a construction project, creating a wedge between the sales price of a house and the cost of buying the land and building the structure. The difference between the dark and light lines in Figure 19.1 measures the difference between house prices and the cost of materials and labor, with the dark line depicting the evolution of real house prices based on CoreLogic’s repeat-sales price index.² Even after the recent collapse of house prices, this index was still about 60% higher in the 2000s than it was in the early 1980s.

¹ Figure 19.1 plots index values. In levels, real construction costs for an economy-quality home are about \$85 per sq. ft. Costs are roughly twice that amount for the highest-quality home for which the data provider, the RSMeans, reports consistent information. Hence, the structure for a modest-quality 2000 sq. ft home costs about \$175,000 to construct and about \$350,000 for a higher-quality version.

² The deflator for this series is that of consumption expenditures excluding housing.

The growing wedge between house prices and construction costs illustrates that the price of land has been trending upward over time. Although a portion of the increase in land prices undoubtedly owes to geographic constraints (a topic to which we will return later), man-made constraints also must play an important role. Otherwise, valuable land would be built out at a far-higher density than currently exists (Glaeser and Gyourko, 2003).³

Interest in how local land use regulation might have influenced the elasticity of housing supply has increased over the past few decades. This attention is at least partly due to a suspicion that the local residential land use regulatory environment has grown stricter and become more binding over time, particularly in areas facing strong demand for entry. Gyourko et al. (2013) reported evidence consistent with the hypothesis that new housing construction in local markets generally was not constrained prior to the 1970s, as during that period, high house price growth almost always was accompanied by a large amount of residential construction. In later decades, house price increases coincided with much less construction, pointing to constraints on the housing supply. The idea that supply constraints began to bite after the 1970s also is consistent with Frieden (1979), who was among the first to argue that what we now call NIMBYism was one factor behind the rise of environmental impact rules to slow or stop development.

Modern-day land use regulation in the United States began with zoning laws in the 1910s with the intention of separating different types of land use, thereby limiting the negative externalities associated with certain types of industrial or commercial use (Fischel, 2004; Quigley and Rosenthal, 2005). By and large, land use in the United States is controlled by local governments. The US Constitution did not grant the federal government authority to regulate land, and the states have generally left this power with local governments.⁴ The fact that land use is controlled by local governments has contributed to the heterogeneity of regulations. Over time, the types of regulations have expanded and now include urban growth boundaries, minimum lot sizes, density restrictions, and height restrictions, among many others.

In this chapter, we review the research on regulation to date and highlight a number of promising areas for future exploration. We restrict our discussion to rules and laws imposed by any form of government that restricts the number, location, quality, or shape of residential development. We do not discuss actions by private organizations such as homeowners associations or regulations pertaining to commercial development because those topics would require a much longer review.

Despite the growing body of literature on housing supply regulation, there is still much that we do not understand. As we discuss more fully in the next section, research on this topic has been hampered by a lack of direct evidence on regulation. The degree of

³ Not all man-made constraints are regulatory in nature. For example, Brooks and Lutz (2012) showed that in urban Los Angeles, frictions related to assembling contiguous parcels of land restrict development.

⁴ Some European countries like the United Kingdom and France do have national planning guidelines.

local land use restrictiveness is challenging to define because constraints can come in so many different forms. The best information on regulation comes from surveys. Some surveys collect data from a large number of jurisdictions within a single metropolitan area, while others collect data from a few jurisdictions in a large number of metropolitan areas. Either way, it is then up to the researcher to combine the results in order to depict most accurately the strength of regulation. Not surprisingly, this is a difficult task because the relative importance of different types of regulations is not well understood. A second major empirical problem is that we do not have good time series with which to measure changes in regulation. With only cross-sectional evidence, it is very difficult to disentangle the causes and effects of regulation from local demographic and socioeconomic characteristics that might be correlated with regulation. A third challenge is that many of the predicted determinants of regulation are likely to have independent effects on housing market outcomes even in the absence of regulation. Thus, even with the benefit of time series data, it is challenging to identify the effects of regulation.

After discussing issues related to the measurement of regulation and reviewing a number of data collection efforts, we turn to research on the determinants of regulation. Many theories have modeled the role of homeowners as the primary supporters of these regulations because they have a clear incentive to block development in order to protect the value of their property. We also discuss the effect that a limited supply of vacant land, stemming from either geographic constraints or past development, may have in encouraging regulation. A third important strand of the literature considers the fiscal and exclusionary motives for zoning.

We then examine the consequences of regulation, the vast majority of which have focused on effects in the housing market. The simplest models predict that regulation will reduce the elasticity of housing supply, resulting in larger house price increases and slower growth in the quantity of housing as demand increases. Measurement issues notwithstanding, most papers do find a strong positive relationship between regulation and house prices and a strong negative relationship between regulation and construction. Regulation also appears to reduce the responsiveness of the housing supply to demand shocks, as well as influence the size of metropolitan areas and the type of structures that are built. There has been much less research examining the effects of regulation beyond housing markets. A few papers have found that regulation is associated with household sorting by income or other demographic characteristics, while others have found that regulation reduces the elasticity of labor supply by altering the migration patterns of workers.

The penultimate section reviews work on the welfare consequences of local land use regulation. Government intervention can have both costs and benefits, so one cannot presume that regulation in this area is inefficient *per se*. Some regulations such as building codes banning asbestos in insulation materials or requiring fire-retardant roofing products almost certainly have benefits that exceed their costs, and consequently, they seem favored by most of society. More general zoning regulations are challenging to evaluate.

For example, [Hamilton \(1978\)](#) showed conceptually that zoning could help local jurisdictions provide efficient levels of public services and allow homebuilders to sort so that housing consumption was efficient, too. However, [Barseghyan and Coate \(2013\)](#) recently showed that this conclusion does not hold in a dynamic context in which existing structures are exempted from any new zoning regulation. In that case, it is possible for there to be “overzoning” in which housing consumption is inefficiently high (although public service provision need not be). Much remains to be done on this important topic, but recent empirically oriented research suggests that the overall efficiency losses from binding constraints on residential development could be quite large ([Glaeser et al., 2005](#); [Turner et al., 2014](#)). In the final section of this chapter, we discuss some areas where more research would be particularly beneficial.

19.2. DATA: OLD AND NEW

Data collection and measurement have lagged behind theory and modeling from the beginning. Perhaps, the most important reason is that development can be affected in a myriad of ways. The heterogeneity of regulations is well illustrated by one of the earliest surveys of land use regulation, a survey of hundreds of communities in California by [Glickfeld and Levine \(1992\)](#) that collected information on 14 different controls and regulations and two catch-all “other” categories. They found that interventions were wide-ranging in nature and included infrastructure requirements (which are called exactions in some more recent surveys), height restrictions, caps on the number of housing units built or permitted, population growth limits, urban boundaries or green zones, restrictions on rezoning for a less intense land use, restrictions on upzoning to a more intense land use (including requiring voter approval), and supermajority rules for a zoning board or city council to approve zoning requests.⁵ Subsequent researchers have asked about additional regulations such as minimum lot size requirements, as well as delays in local government decision-making (often called “approval lags”).

There is no agreement among scholars or practitioners upon a single definition of the degree by which development in a local land market is constrained. Nor is there a consensus that one particular regulation or subset of constraints is most important. In the absence of a common understanding, it should not be surprising that empirical work has focused on aggregate measures of regulation, often indexes of individual rules. If communities use different types of regulation as substitutes, then measuring only one type could provide a misleading picture of the locations where regulation is generally stricter. But it is difficult to collect accurate data on the wide variety of regulations in place, not to mention to compare the stringency of one type of regulation with another.

⁵ The typical jurisdiction in the [Glickfeld and Levine \(1992\)](#) survey had at least two of the above interventions. All types of regulation existed in multiple jurisdictions. See table 2 of their paper for more details.

19.2.1 Indirect measurement

One way to bypass the difficulty of accurately measuring complex local land use environments is to impute the presence of binding restrictions using simple economic analysis. Regardless of whether binding local constraints on housing supply exist, standard neoclassical theory implies that price equals average cost as long as the market is sufficiently competitive. Glaeser et al. (2005) argued that a large gap between prices and marginal construction costs is a clear signal that a market is tightly regulated. It is worth noting that the housing supply can be slow to adjust to changes in demand, so prices can deviate from costs in the short run even in an unregulated market. Thus, the fact that the dramatic rise and fall of house prices in the late 2000s does not appear to be correlated with observable measures of regulation, as observed by Davidoff (2013), does not conflict with the idea that *persistent* deviations of prices from costs reflects regulation.

Glaeser et al. (2005) applied their argument to the market for Manhattan condominiums, where in the absence of regulation, building heights should rise to the point where the marginal cost of adding an extra floor equals average cost. Regardless of restrictions on building size, free entry of builders still should keep price equal to average cost. Assuming that the marginal cost function is increasing, as seems plausible because it is more challenging to build up, regulation implies that both prices and average costs will be above marginal costs. Consequently, Glaeser et al. (2005) used the difference between market price and the marginal cost of building an additional floor to proxy for the extent of housing supply restrictions.⁶ They labeled this gap a “regulatory tax” and concluded that it caused Manhattan condominium prices in the beginning of the 2000s to be at least 50% higher than would the case under a completely free development policy.

Glaeser and Gyourko (2003) and Glaeser et al. (2005) performed a similar analysis for single-family housing markets across the nation. While the underlying economic logic is the same, empirical implementation is more difficult for this sector because one needs to know the free market price of land in order to determine the marginal cost of supplying an extra home. Unfortunately, there are relatively few observations on sales of vacant land, especially in densely occupied areas.⁷ Various statistical techniques involving hedonics can be used, but estimation error of the regulatory tax is likely to be larger in the single-family sector than in a market dominated by high rises.⁸ That caveat aside, Glaeser et al. (2005) reported a wide range of “regulatory tax” estimates for single-family

⁶ Consulting engineers can provide fairly precise estimates of this cost because additional land is not needed to build up.

⁷ Recently, some researchers have used data on land sales collected by the CoStar Group, Inc. See, for example, Turner et al. (2014) and Nichols et al. (2013).

⁸ The recent house price cycle provides a helpful illustration of how incorporating land prices matters for calculating the regulatory tax. Davidoff (2013) showed that the magnitude of the recent boom and bust in house prices across metropolitan areas is uncorrelated with the degree of regulation. However, one should not conclude that the regulatory tax is uncorrelated with regulation because land prices rose and fell sharply in the same areas where the house price cycle was most pronounced (Nichols et al., 2013).

markets across the country as of 1999–2000, ranging from zero in Birmingham, Cincinnati, and Houston to nearly 20% of total house value in Boston, over 30% in Los Angeles, and upward of 50% in the San Francisco Bay Area.

Not only does one not need detailed information on the local regulatory environment with this approach, but also it provides a natural way to summarize the strength of regulations across a heterogeneous set of laws that affect land use. The downside, of course, is that the evidence is indirect, so it cannot provide insight into the impact of any specific policy. Moreover, for most housing markets, it requires an assessment of the marginal price of vacant land, which is not easy to measure. Given the difficulty in measuring land prices and the dependence of this approach on certain assumptions, it seems wise to conclude that local regulatory restrictions are binding and economically important only when the estimated “regulatory tax” is large.⁹ Even so, the available research suggests this tax is quite large for many markets, which indicates that it is worthwhile for researchers to invest in examining specific regulations in different locations across the nation.

19.2.2 Building codes

Among the first empirical efforts to collect data on housing supply regulation were those that looked into the role of building codes, which are regulations with the stated purpose of assuring the safety and sanitation of homes and limiting the negative externalities associated with low-quality construction. These regulations tend to raise the cost of construction by restricting the types of materials used or the method of construction, but they do not directly limit the quantity of housing that can be built on any given amount of land.

Noam (1983) is an early example of research on this topic, which used data collected by the City Managers’ Association for 1100+ cities and towns to create an index of strictness of the local building code based on 14 traits or code provisions.¹⁰ We discuss empirical studies of the impact of regulation in Section 19.4, but suffice it to say that Noam (1983) reported a statistically significant correlation between local house prices and building code strictness in a cross-sectional regression of price on a code strictness index and other covariates. As noted in the “Introduction,” the higher price well may be worth the cost, so no welfare conclusions can be drawn from this correlation. Moreover, as we will discuss in Sections 19.3 and 19.4, building codes may be more common in high-priced areas so the simple correlation between regulation and house prices likely overstates the true causal effect.

⁹ Housing is heterogeneous in nature, so small differences between prices and costs could reflect measurement error in either variable. See Glaeser et al. (2006a) for a discussion of other relevant assumptions, the most important being a competitive supply of homebuilders.

¹⁰ See Colwell and Kau (1982) and Oster and Quigley (1977) for other early analyses using data on building codes.

While building codes appear to be correlated with housing prices, they likely raise the price of the housing structure rather than the land that the housing unit occupies. As we showed above, it appears to be land prices that are the root of high and rising house prices in the United States. This conclusion is confirmed by other research showing that land prices have been rising for the nation in aggregate for some time, in contrast to real construction costs (e.g., [Davis and Heathcote, 2004](#); [Davis and Palumbo, 2008](#); [Nichols et al., 2013](#)). As land has become a larger share of the total price of a home, it is perhaps not surprising that a growing amount of effort has turned to measuring local land use restrictions rather than the regulations that affect the quality of residential structures.

19.2.3 Land use controls

The first studies of land use regulation generally involved narrowly targeted data collection efforts to analyze a specific issue. In a study of how growth controls affect house price growth, [Segal and Srinivasan \(1985\)](#) asked staff members of local and regional planning agencies across 51 metropolitan areas to estimate the percentage of land removed from development out of otherwise developable suburban land. [Katz and Rosen \(1987\)](#) were even more targeted, using information from a mail and phone survey conducted by the Center for Real Estate and Urban Economics at the University of California–Berkeley on growth management programs in cities and towns throughout the San Francisco Bay Area. A community was deemed to operate under a growth management program if it had a specific cap on the number of building permits that could be issued in any given year between 1973 and 1979. A simple OLS regression of local house prices on a dichotomous indicator for the presence of a growth management program found that prices were 17% and 38% higher in communities with such programs.¹¹ Of course, as we will discuss below, this cross-sectional correlation probably gives a biased estimate of the effect of regulation because communities with growth management programs are different from other communities in many ways that likely affect house prices.

Such an economically large relationship piqued interest in the role of local land use constraints and helped stimulate the development of broader indexes of regulatory control. Among those that followed were [Linneman et al. \(1990\)](#), [Glickfeld and Levine \(1992\)](#), [Levine \(1999\)](#), [Evenson and Wheaton \(2003\)](#), [Foster and Summers \(2005\)](#), [Quigley and Rosenthal \(2005\)](#), [Pendall et al. \(2006\)](#), [Glaeser et al. \(2006b\)](#), [Gyourko et al. \(2008\)](#), [Glaeser and Ward \(2009\)](#), and [Saiz \(2010\)](#). Space limitations prevent us from reviewing each data collection effort in detail. Fortunately, the range

¹¹ Other early data collection efforts include a 1984 survey on economic development conducted by the International City Management Association. This survey was sent to the chief administrative officer of various cities and contained four questions pertaining to constraints on construction. See [Clark and Goetz \(1994\)](#) for more on this survey. At the state level, the American Institute of Certified Planners conducted a survey of the types of land use planning activity undertaken by each of the 50 states. Those results were published in a 1976 book *Survey of State Land Use Planning Activity*.

of efforts, including their strengths and weaknesses, can be illustrated by contrasting three recent empirical efforts: Glaeser et al. (2006b), Gyourko et al. (2008), and Saiz (2010).

The first question faced by data collectors in this area is precisely what to measure. Perhaps, the most straightforward way is to place a numerical cap on new quantity—the type of restriction cataloged and studied by Katz and Rosen (1987). But the supply of housing can be limited in many other ways including minimum lot sizes, height restrictions, setback requirements, and open-space designations. Regulations that affect the cost of construction can also influence the supply of housing, as indicated by Noam's (1983) analysis of local building codes. Even these examples do not come close to capturing the range of efforts devised by communities to influence housing supply. Creativity on the part of local governments appears to know virtually no bounds in this instance.

Heterogeneity in land use restrictions across localities is so extensive that it is almost impossible to describe the full complexity of the local regulatory environment. One strategy is to focus on a small set of locations and collect as much detailed data as possible on all aspects of the regulatory environment in those places. Glaeser, Schuetz, and Ward (GSW, hereafter; 2006b) did so for a subset of the Boston metropolitan area. Their detailed analysis of local zoning provisions allowed them to compute fairly precise estimates of the potential housing supply in all localities in their sample. However, the enormity of that effort prevents it from being easily replicated in many other markets.

Whereas GSW (2006b) reflected a “deep but narrow” approach to studying regulation, an alternative strategy is to go “shallow but wide.” A good example of this latter approach is Gyourko, Saiz, and Summers (GSS, hereafter; 2008), who estimated indexes of the stringency of the local land use environment for 2611 communities across the nation, of which approximately three-quarters are in 293 distinct metropolitan areas spread across all 50 states. They achieved this breadth of coverage at the cost of detailed knowledge of the environment.

GSS (2008) constructed their regulatory index based on the answers to three sets of questions. The first inquired about the general characteristics of the regulatory process: Who is involved in that process (e.g., states, localities, councils, legislatures, and courts) and who can approve or veto zoning or rezoning requests. A second set of questions pertained to the local rules: Were there binding limits on quantities supplied? Minimum lot size requirements? Affordable housing requirements? Development exactions of various types? The final set of survey questions asked about changes in the cost of lot development and in the review time for a typical project over the previous 10 years.¹²

¹² GSS (2008) supplemented their survey responses with data from two other sources. One was the state-level analysis of the nature of legal, legislative, and executive actions pertaining to land use policy that was conducted by Foster and Summers (2005). The other was a measure of community pressure using information on environmental and open-space ballot initiatives. The interested reader should see GSS (2008) for more on these data.

The answers to these questions were used to create an aggregate measure of the stringency of the local land use regime that GSS called the Wharton Residential Land Use Regulation Index (WRLURI). Specifically, they used factor analysis to create this summary metric from 11 subindexes that described different aspects of the local regulatory environment, with the results standardized to have a sample mean of zero and a standard deviation of one.¹³ The index is increasing in the restrictiveness of regulation. Table 19.1 reproduces table 11 from GSS (2008) that reports average WRLURI values for the 47 metropolitan areas in their sample that had more than 10 communities responding to the survey.

The index itself is constructed so as to rank places in terms of the degree of strictness of the local residential land use environment, but the underlying survey data allow us to describe an average community, as well as strongly and weakly regulated ones. The typical community in the WRLURI sample can be characterized as follows: (1) two entities (zoning commission, city council, environmental review board, etc.) must approve any project requiring a zoning change; thus, multiple points of approval now are required for projects that cannot be done “by right”¹⁴; (2) there is a modest minimum lot size requirement, but for the typical community, it is far less onerous than the 1 acre minimums seen in highly regulated places; (3) some type of development exaction program exists¹⁵; and (4) there is about a 6-month lag between the submission of a permit for a standard project (where standard is defined by the community filling out the survey) and the issuance of that permit.

Dividing the WRLURI communities into thirds and labeling the top third “highly regulated” and the bottom third “lightly regulated” allows us to contrast traits across those two groups. One noteworthy distinction is that local and state pressure groups are much more likely to be involved in the regulatory process in the more highly regulated places. Another key distinguishing feature between highly and lightly regulated communities is that more than 50% of the highly regulated communities have a 1 acre minimum lot size rule in at least one neighborhood, whereas less than 5% of lightly regulated communities have such a rule. Open-space requirements and some type of formal development exactions policy are nearly omnipresent in the highly regulated localities, but often are not present in the most lightly regulated areas. In addition, the average delay time between

¹³ More specifically, the principal component of each subindex was used to create the community-wide index. The 11 subindexes include a local political pressure index, state political involvement index, state court involvement index, local zoning approval index, local project approval index, local assembly index, supply restrictions index, density restrictions index, open-space index, exactions index, and approval delay index. The literature has not reflected on how to aggregate across individual regulations, but it certainly would be beneficial to consider what an optimal weighting scheme might look like.

¹⁴ A project can be done “by right” if it meets all published zoning regulations and requires no variances of any kind.

¹⁵ Development exactions are monetary or in-kind payments in return for development rights.

Table 19.1 Average WRLURI values for metropolitan areas with data on 10 or more local jurisdictions

Metropolitan area	WRLURI	Number of observations	Metropolitan area	WRLURI	Number of observations
1. Providence–Fall River–Warwick, RI–MA	1.79	16	25. Milwaukee–Waukesha, WI	0.25	21
2. Boston, MA–NH	1.54	41	26. Akron, OH	0.15	11
3. Monmouth–Ocean, NJ	1.21	15	27. Detroit, MI	0.12	46
4. Philadelphia, PA	1.03	55	28. Allentown–Bethlehem–Easton, PA	0.10	14
5. Seattle–Bellevue–Everett, WA	1.01	21	29. Chicago, IL	0.06	95
6. San Francisco, CA	0.90	13	30. Pittsburgh, PA	0.06	44
7. Denver, CO	0.85	13	31. Atlanta, GA	0.04	26
8. Nassau–Suffolk, NY	0.80	14	32. Scranton–Wilkes–Barre–Hazelton, PA	0.03	11
9. Bergen–Passaic, NJ	0.71	21	33. Salt Lake City–Ogden, UT	−0.10	19
10. Fort Lauderdale, FL	0.70	16	34. Grand Rapids–Muskegon–Holland, MI	−0.15	16
11. Phoenix–Mesa, AZ	0.70	18	35. Cleveland–Lorain–Elyria, OH	−0.16	31
12. New York, NY	0.63	19	36. Rochester, NY	−0.17	12
13. Riverside–San Bernardino, CA	0.61	20	37. Tampa–St. Petersburg–Clearwater, FL	−0.17	12
14. Newark, NJ	0.60	25	38. Houston, TX	−0.19	13
15. Springfield, MA	0.58	13	39. San Antonio, TX	−0.24	12
16. Harrisburg–Lebanon–Carlisle, PA	0.55	15	40. Fort Worth–Arlington, TX	−0.27	15
17. Oakland, CA	0.52	12	41. Dallas, TX	−0.35	31
18. Los Angeles–Long Beach, CA	0.51	32	42. Oklahoma City, OK	−0.41	12
19. Hartford, CT	0.50	28	43. Dayton–Springfield, OH	−0.50	17
20. San Diego, CA	0.48	11	44. Cincinnati, OH–KY–IN	−0.56	27
21. Orange County, CA	0.39	14	45. St. Louis, MO–IL	−0.72	27
22. Minneapolis–St. Paul, MN–WI	0.34	48	46. Indianapolis, IN	−0.76	12
23. Washington, DC–MD–VA–WV	0.33	12	47. Kansas City, MO–KS	−0.80	29
24. Portland–Vancouver, OR–WA	0.29	20			

Notes: This table appears as table 11 in [Gyourko et al. \(2008\)](#). Metropolitan area definitions are based on the 1999 boundaries. Consolidated metropolitan statistical areas are disaggregated into primary metropolitan statistical areas wherever relevant.

application and approval for a standard project in the highly regulated communities is three times longer than in the lightly regulated communities (10.2 months vs. 3.2 months, respectively).

Another noteworthy feature of the WRLURI is the strong positive correlation across its component indexes. Essentially, if a community is rated as highly regulated on one dimension, it is likely to be highly regulated along other dimensions. This result suggests that surveys covering only a limited set of types of regulation might provide a reasonably accurate picture of the general restrictiveness of the locations in the sample, but further research should examine this conjecture more thoroughly.

In terms of data detail and market coverage, the contrast of GSS (2008) with GSW (2006b) is stark. Whereas GSS (2008) only know whether a town had a 1 acre minimum lot size restriction in at least one neighborhood, GSW (2006b) know how every square yard in the entire municipality is zoned. For example, there are 14 municipalities in the GSW sample with minimum lot sizes in excess of 70,000 sq. ft (over 1.6 acres). Those areas constitute more than 10% of the region's land area but hold only 4% of the population. More generally, the diversity of minimum lot size restrictions in the Boston area is staggering, with the range running from less than 10,000 sq. ft to more than 70,000 sq. ft. The modal minimum lot size restriction is roughly 1 acre, providing some indication of how stringent the zoning codes are in many of the communities in the GSW sample.

The detailed knowledge of the local regulatory environment amassed by GSW (2006b) was collected over a 2-year period in which the authors worked with a team of researchers at the Pioneer Institute for Public Policy Research to catalog the land use regulatory environment in 187 communities that comprise a subset of the Boston metropolitan area. They conducted a survey of over 100 questions, interviewed local officials, and provided those individuals the opportunity to review the results. Such detail is only possible within a given metropolitan area and could only be replicated in other regions at high cost.

The benefits are considerable. The Boston metropolitan area is worthy of detailed analysis because key summary statistics suggest it is very tightly regulated. In the midst of high and rising real house prices over time, the number of housing permits has shrunk considerably: from 172,459 in the 1960s to 141,347 in the 1980s and down to 84,105 in the 1990s (GSW, 2006b). GSW (2006b) recognized the possibility that the reason for the downward trend in construction could be that Boston is running out of land, rather than man-made regulation. However, the authors showed that densities outside the urban core are quite low, suggesting that land is still plentiful in the Boston area.¹⁶

¹⁶ Moreover, if it were land that really was scarce, the price of a quarter acre of extra land would be the same whether it extends an existing lot or sits under a new home. Glaeser and Gyourko (2003) showed this not to be the case for the Boston area.

This result led GSW (2006b) to a detailed study of local zoning codes. In doing so, they provided a rare window into the intricacy of these laws. Their survey includes not only the hard caps on growth akin to those studied by Katz and Rosen (1987) but also the prohibition of irregularly shaped lots, extensive wetlands restrictions, septic system regulations, and various subdivision rules. Not all rules are designed to limit supply, as there are also “cluster provisions” and inclusionary zoning provisions, which allow developers to build at higher densities in some places, and age-restricted zoning that typically encourages building for seniors.

GSW (2006b) used historical data on permits in their communities (going back to 1910 using decennial Census data) to investigate which of the myriad of regulations really matter in terms of constraining residential construction. They concluded that evidence linking minimum lot size restrictions to new development is especially persuasive. As minimum lot size increased by a quarter acre, roughly 10% fewer houses were permitted over time. This negative relationship between regulation and permits was found even in communities with lower minimums. Relatively strict wetland regulation also appears to constrain new development in an economically meaningful way. In a follow-up paper (Glaeser and Ward, 2009), two of these authors used these data to examine the causes and consequences of regulation in more detail. We will discuss those findings below.

19.2.4 Other measures

Beyond investigating building codes or land use regulation, another method of inferring restrictiveness of supply can be found in Saiz (2010), who incorporated the role of topography to estimate the elasticity of housing supply for nearly 100 metropolitan areas in the United States. His approach used three separate data sources. First, he employed GIS techniques to compute the area lost to oceans within a 50 km radius of the centroid of a metropolitan area. Second, he used satellite data from the US Geological Survey (USGS) to calculate the amount of land lost to internal water bodies and wetlands. Third, under the rationale that anything above a 15° slope is extremely challenging and expensive to build on, he used the USGS Digital Elevation Model to compute the percentage of land area with a slope in excess of 15°. Taken together, these three computations provide an estimate of exogenously undevelopable land in a metropolitan area.

Saiz (2010) then proceeded to investigate links between geography, regulation, and urban development. As we will discuss below, he found a strong statistical link between his estimates of geographic restrictions and the WRLURI. That is, regulatory restrictiveness tends to be higher in metropolitan areas that have geographic constraints on land development in terms of being coastal, having internal water bodies of some type and/or steep elevations. Based on the geographic constraints and WRLURI, Saiz (2010) estimated the elasticity of housing supply for each of the metropolitan areas in

his sample. These estimates have been used by many other researchers to analyze the housing supply and to provide exogenous variation in house prices across locations.

19.2.5 Summary of measurement issues and concluding comments

In the past decade or so, there have been considerable achievements in the measurement of local land use regulations. Early studies focused on only a single type of regulation within a fairly limited geographic area. By contrast, GSW (2006b) provided a very detailed picture of the entire regulatory environment of a single metropolitan area. While it is very costly to amass this type of information, it would be helpful to replicate their survey in other metropolitan areas as often as possible. Combining data from such detailed studies of land use regulation would be even more valuable than examining a single area in isolation.

More effort has been put into less specific and more general measurement of local land use regulatory environments across a wider range of communities, as reflected in the GSS (2008) survey of more than 1900 communities in over 290 metropolitan areas. Although this survey does not contain detailed information on all the different ways local governments can constrain residential development, the advantage is in allowing researchers to compare regulation across many more localities and markets.

Beyond surveys, Saiz (2010) illustrated the possibility of using other data to gain insight into the elasticity of housing supply. Using the geography and topography of the land, he documented a connection between geographic constraints and regulatory constraints and provided estimates of housing supply elasticities for nearly 100 metropolitan areas.

Because surveys are so expensive in terms of time and money, researchers should be creative in thinking about other ways to measure regulation using readily available data. One possibility is to use public records data, which are now routinely collected and sold by a few data providers. These datasets, which cover the vast majority of residential property in the United States, include parcel-by-parcel information on allowed land use. For example, Brooks and Lutz (2014) use land use codes from public records data to analyze zoning and public transit in Los Angeles. Another possibility might be to use data from Internet chatter, as in Gyourko and Saiz (2006).

As will be documented below, measures of local land use restrictiveness have been widely used in subsequent empirical research into the causes and consequences of regulation. However, they are limited in the sense that each is a cross-section. Because regulation can be correlated with many other attributes of a location, a single snapshot of the local regulatory environment is not conducive to establishing causality in empirical work. Thus, it is critical that researchers collect new data or otherwise come up with ways to measure changes in regulation over time. For example, Glaeser and Ward (2009) converted the GSW (2006b) survey into panel data based on the year that different types of

regulation are adopted. Similarly, Jackson (2014) converts the Glickfeld and Levine (1992) survey into a panel of regulations in California.¹⁷ Besides panel data, another helpful approach would be to examine changes in a particular regulation using a difference-in-difference strategy. A few examples are Zhou et al. (2008), who analyzed the effect of an amendment to the Chicago Zoning Ordinance; Cunningham (2007), who analyzed the adoption of an urban growth boundary in Seattle; Kahn et al. (2010), who analyzed the creation of a coastal boundary zone in California; and Thorson (1997), who analyzed an increase in minimum lot sizes in McHenry County, IL.

19.3. DETERMINANTS OF REGULATION

19.3.1 Joint determination of land values and zoning

One of the earliest questions asked in the literature on the determinants of zoning is whether local governments set zoning laws in order to maximize the value of land in their jurisdiction. For example, Wallace (1988) modeled zoning as the outcome of a county council that maximizes its utility, which is a function of land attributes and land prices. Land prices are also a function of local characteristics, as well as of the zoning designation chosen. If zoning has no effect on land values, then one would conclude that these regulations merely coincide with the land use that would be adopted in unconstrained markets. Wallace refers to this question as whether “zoning follows the market.” In a sample of parcels for Kings County, WA, she found that zoning allocations tend to decrease the price of parcels zoned for large lots, suggesting that zoning leads to an oversupply of large minimum lot sizes.¹⁸ More generally, she found that land values would be higher under different zoning designations. Thus, she concluded that in this sample, zoning does not “follow the market.” Her analysis uses observed attributes of land parcels to correct for the selection of different types of land into different zoning categories. However, it is difficult to rule out the alternative interpretation that unobserved variables cause cheaper land to be zoned for larger lots.

McMillen and McDonald (1991) also estimated a model in which land values affect zoning decisions, and vice versa. In their data from the northwest suburbs of Chicago, they identified the determinants of zoning from an indicator of whether a parcel was located in Cook County, because that county was more likely to zone undeveloped land

¹⁷ Although these efforts are clearly a vast improvement over cross-sectional data, they do not account for regulations that no longer existed at the time of the survey. Thus, if cost were not an issue, repeated surveys would be a better method of constructing panel data than using the implementation date of existing regulations.

¹⁸ Because zoning and land values are jointly determined, she identifies a two-equation model of land values and zoning using an indicator for whether a parcel is platted as an instrument for land values. The justification for this specification is that approval of a plat application entails considerable time and monetary costs, raising the price of land, and that the zoning designation is decided prior to when a parcel is platted.

for residential use. They found that zoning patterns reflect characteristics of the land in that parcels closer to railways and expressways are more likely to be designated for manufacturing use. They also showed that estimates of the land value equation are biased when one does not account for the endogenous designation of zoning.

In a sample of home sales in Santa Clara County, CA, [Pogodzinski and Sass \(1994\)](#) differentiated between the effects of land use designations and regulations that affect the characteristics of a lot and its structure. They found that land use designations do not affect house values, implying that this type of zoning does “follow the market.” By contrast, minimum lot sizes, minimum side-yard restrictions, and maximum height restrictions do affect house values, implying that these types of regulations do not “follow the market.”

To summarize, this early work has generally found that zoning regulations do not simply mimic the outcome that would result from an unregulated market. However, because their focus is on the effect of regulation on land values, these papers provide little insight into the factors that do shape housing supply regulation.

19.3.2 Homeowners, developers, and local politics in a single community

More recent theories concerning the determinants of regulation have incorporated the idea that local residents do not all share the same goals, so the regulatory environment will be shaped by the incentives and influence of actors in the local political process. In a number of influential articles and books, William Fischel has advanced the idea that homeowners are an important force behind these types of regulation.¹⁹ In *The Homevoter Hypothesis* (2001), he contended that homeowners have a strong incentive to work together in order to restrict undesired development because their home is typically their primary asset and owners do not have other means to insure against events that would reduce their property value. Central to Fischel’s argument is the idea that local amenities and disamenities (such as school quality and dumps) are capitalized into house values. In support of this notion, he cited a number of studies finding that house prices are lower in neighborhoods with disamenities such as a toxic waste dump, more through traffic, or more localized air pollution, as well as research showing that property tax rates are capitalized in house prices. Because shocks to housing wealth are difficult to insure or diversify, homeowners turn to local political action to protect their investment.

According to Fischel, another reason why homeowners are so opposed to development is that they tend to work in a different jurisdiction than where they live. When construction is pushed into a nearby jurisdiction, homeowners can benefit from some

¹⁹ Fischel drew on earlier work discussing the role of homeowners in constraining local development including [Sonstelie and Portney \(1978\)](#) and [Ellickson \(1977\)](#). In an early book, Fischel proposed that local governments impose land use regulation in order to maximize the value of owner-occupied housing ([Fischel, 1985](#)).

positive effects, such as job creation, without incurring the cost of higher density in their own neighborhood. This logic can explain both why regions with more fragmented governments are more likely to have stricter regulation (Fischel, 2008) and why regulation spread and became stricter over the twentieth century along with suburbanization (Fischel, 2004).

Drawing on Fischel's insights, a number of papers have developed more formal theoretical models of homeowners' influence on zoning regulation. Ortalo-Magne and Prat (2014) present a cutting-edge example, in which they used an overlapping-generations' economy to model three decisions simultaneously: the location choice of households (a regulated city versus unregulated countryside), investment in residential real estate, and the collective choice model of local housing supply regulation.

Their basic setup is as follows. Households decide whether to live in the countryside, where they pay a fixed rent and earn income normalized to zero, or in a city where income is a function of citywide productivity that varies over time and a fixed idiosyncratic component. In the city, households choose whether to invest in owner-occupied housing or to rent. There is a fixed premium between the return to investing in owner-occupied housing and the return to investing in rental housing, which is meant to reflect a number of factors including the nonpecuniary benefits of homeownership and its favorable tax treatment, the costs of managing rental property, and moral hazard in the rental market. Housing wealth is consumed at the end of an individual's life. Thus, tenure choice is a trade-off between the homeownership premium and the price risk associated with homeownership.²⁰ New construction in the city requires a building permit, and the number of building permits allowed is decided by city residents through majority voting. Permits are assigned to developers, who do not vote and who must pay a fee for every new permit. Aggregate permit fees paid are split equally among all city residents. These fees can be interpreted as a rough proxy for the general economic gains that development can bring to an area.

Not only do city residents receive a share of the total permit fees raised, but also they benefit from new construction because the larger housing supply reduces future rents and all city residents must pay rent either to themselves or to absentee landlords. However, the reduction in rents also causes house prices to fall, leading to lower end-of-period consumption. In the model, construction does not occur if the net benefit accruing to the median voter is negative—i.e., if the benefit owing to lower future rents and the voter's share of permit fees is less than the decrease in final consumption through lower house prices.

Using this model, Ortalo-Magne and Prat showed that a homeownership subsidy makes homes less affordable. Absent the constraints on new construction, the subsidy

²⁰ Since households are risk-averse, without the homeownership premium, everyone would choose to rent. Housing not owned by households is owned by risk-neutral firms.

would have no effect on affordability because it would be exactly offset by an increase in house prices. However, the subsidy causes the median city resident to own more housing, thereby leading to more opposition to city growth. It is therefore the reduction in permitted new construction that raises equilibrium house prices and reduces affordability.

In contrast to the homeownership subsidy, an increase in the permit fee boosts support for urban growth, leading to an increase in city size. The expansion of the housing supply reduces rents, making all city residents better off. This result illustrates Fischel's argument that the more homeowners internalize the benefit from new construction, the less opposed to development they will be.

Theories in the vein of Fischel and Ortalo-Magne and Prat predict that homeowners should be stronger supporters of housing supply regulation than renters; other examples include Glaeser et al. (2006b) and Hilber and Robert-Nicoud (2013). A few studies have found support for this hypothesis in voting patterns. For example, Dubin et al. (1992) documented that precincts with a larger share of homeowners had a larger proportion of votes cast in favor of growth controls in a 1988 San Diego election. McDonald (1995) also examined voting patterns—in this case, a proposed zoning ordinance in Houston—and found that middle-income precincts were more likely to support a proposed zoning ordinance than lower-income precincts. Although he does not examine homeownership directly, it is likely that middle-income precincts have higher homeownership rates than lower-income precincts. An interesting aspect of McDonald's study is that zoning did not exist in Houston, so the voting patterns do not reflect any preexisting pattern of regulation.

If homeowners favor tighter regulation, then policy should be more restrictive in areas where the political influence of homeowners is stronger. However, evidence for this relationship is difficult to find. Logan and Zhou (1990) found little correlation between homeownership and a variety of growth control measures in a national sample of suburban municipalities in 1973. Similar results have been found for growth control measures in cities around San Jose (Baldassare and Protash, 1982) and for municipalities in California (Donovan and Neiman, 1995; Brueckner, 1998). It is possible that the mere presence of homeowners is not a particularly accurate measure of the influence that these constituents have on local policy. Instead, their influence might be better reflected by demographic or socioeconomic characteristics of the local population that are correlated with their ability to organize or otherwise participate in local policymaking. Many studies have found positive correlations of regulation with socioeconomic characteristics that are likely to be correlated with the political influence of homeowners, such as income, education, and the fraction of white-collar or professional workers.²¹ However, it should be noted that because all of these studies are cross-sectional in nature, it is difficult to rule out

²¹ See Lenon et al. (1996), McDonald and McMillen (2004), Evenson and Wheaton (2003), Donovan and Neiman (1995), and Baldassare and Protash (1982). Glickfeld and Levine (1992) is an exception.

the possibility that the correlations (or lack thereof) are driven by omitted variables. Moreover, the model proposed by Ortalo-Magne and Prat demonstrates that because regulation raises house prices, the degree of regulation could influence the incentive to be a homeowner, and this reverse causality could further obscure these cross-sectional relationships.

Relatively few papers have attempted to account for potential endogeneity problems when examining the role of homeowners in restricting the housing supply. Hilber and Robert-Nicoud (2013) instrumented for homeownership with the fraction of households that are married couples with no children. The idea is that married couples will have higher and more stable incomes than single adults or households with children, making them more likely to qualify for a mortgage. Also, married couples are more likely to be in stable relationships compared with unmarried couples, implying a lower probability of moving and therefore a longer period over which to recoup the transaction costs of purchasing a home. The use of this instrument requires the assumption that the fraction of households that are married couples with no children does not affect regulation through any channel except the homeownership rate. In a sample of 93 metropolitan areas in 1990, they found that locations with a larger share of married households with no children had a higher homeownership rate in the same year. Using this instrument, they found a positive, but insignificant, correlation with the WRLURI described in the previous section. Moreover, the estimated effect is not large, as a one standard deviation increase in a MSA's homeownership rate is associated with a one-third standard deviation increase in regulatory constraints. Of course, one still might be concerned that the fraction of households that are married and without children is either influenced by housing supply regulation or correlated with omitted variables that are also correlated with regulation.

Glaeser and Ward (2009) used a different method to get around the endogeneity of homeownership and regulation, examining relatively recent patterns of housing supply regulation as a function of homeownership rates in 1940 or 1970. This method avoids the problem of reverse causality, but it is less clear that it mitigates concerns about omitted variables, since historical homeownership rates could also be correlated with many unobserved attributes that affect regulation. In their sample of townships in the Greater Boston area, the authors find little correlation of historical homeownership with a number of measures of regulation such as minimum lot sizes, wetland bylaws, septic rules, and cluster zoning. They also found little correlation of regulation with other 1970 attributes of townships that might be correlated with homeowner influence, such as educational attainment and the fraction of the population that is foreign born.

In summary, there is little empirical evidence that areas with more homeowners adopt stricter housing supply regulations, a fairly surprising result given the many theories that point to homeowners as the drivers of regulation. Because much of this evidence is cross-sectional in nature, more work is needed to address the issues of omitted variables and

reverse causality. Panel data would add a useful dimension, allowing researchers to observe how regulation changes after homeownership patterns change. As noted above, except in a very few cases, panel data do not yet exist. Hence, it is vital to explore other instruments for homeownership. One possibility could arise from regulatory changes in mortgage markets that have changed the ease of getting a mortgage in a way that has had differential effects on homeownership across locations, but would otherwise be uncorrelated with factors that affect land use policy. There are multiple conditional clauses in that sentence, so it is by no means certain that such an instrument would be found to be powerful and valid. However, it is critical for empirical researchers to think more creatively along these lines if the causal effect of homeownership on regulation is ever to be uncovered.

Besides homeowners, developers and/or owners of vacant land may also be important participants in the local political process. [Molotch \(1976\)](#) coined the phrase “growth machine” in reference to the idea that landowners in a community will band together in support of policies that will fuel local economic growth. He argued that the people who end up in local government tend to be local business leaders and other types that are naturally sympathetic towards local growth. Thus, local policy can end up favoring policies that promote economic activity, even if the typical citizen does not. One possible explanation for the lack of cross-sectional correlation between homeownership and regulation is that areas with higher homeownership rates may also have stronger alliances favoring progrowth policies, perhaps because the population is wealthier. [Fischel \(2008\)](#) argued that developers have more influence in large governments, causing metropolitan areas with more political fragmentation to have stricter regulation.

Few formal models have explicitly allowed for the role of developers in shaping housing supply regulation. [Hilber and Robert-Nicoud \(2013\)](#) and [Glaeser et al. \(2005\)](#) showed how local planning boards can be influenced by both homeowners and developers. Like Ortalo-Magne and Prat, these models predict that locations with a larger share of homeowners will have stricter housing supply regulation. In addition, they illustrated that regulations will be tighter in jurisdictions where homeowners have a stronger political influence relative to developers. However, the way in which they model this influence is fairly simple. In [Glaeser et al. \(2005\)](#), homeowners use time to influence the local planning board whereas developers use cash, with the result that regulation will be tighter in areas where the planning board is more influenced by time. In [Hilber and Robert-Nicoud \(2013\)](#), owners of developed land (homeowners) and owners of undeveloped land (developers) both use cash to influence the planning board, and their lobbying contributions are a positive function of the value of their land.

This topic is yet another case in which theory is ahead of empirics, as we are aware of little empirical evidence on the influence developers have in shaping housing supply regulation. One study with indirect evidence on this front is [Solé-Ollé and Viladecans-Marsal \(2012\)](#), who argued that developers should be less successful at influencing local

regulations when the currently elected officials have won the vote by a narrow margin. In that case, the elected officials will place a higher priority on following voters' preferences in order to be reelected. Using data on more than 2000 Spanish municipalities from 2003 to 2007, they reported evidence in support of this hypothesis.

19.3.3 Supply of buildable land and historical density patterns

One interesting aspect of the model proposed by Hilber and Robert-Nicoud (2013) is that the incentives of homeowners and developers to influence the planning process are a function of the value of land. Thus, anything that raises land values will affect regulation. Specifically, in their model, jurisdictions have innate differences in amenities and households choose among jurisdictions. The greater demand to live in an area with more valuable amenities is only partly offset by congestion costs, leading to a positive correlation between amenities and development across locations. In equilibrium, locations with more desirable amenities will have a larger share of developed land and, consequently, more land use regulation in order to prevent further development from raising congestion disamenities even more. Consequently, a restricted supply of land (higher share of developed land) causes further restrictions on supply in the form of government regulation.

In support of the prediction that locations with more developed land will have more regulation, Hilbert and Robert-Nicoud showed that metropolitan areas with a larger fraction of developed residential land area (measured using the 1992 National Land Cover Data) have a larger degree of regulation according to the WRLURI. This correlation exists even after controlling for the homeownership rate, so it does not simply reflect the fact that homeowners have a greater influence on local policy in areas where they are more numerous. Further, Hilber and Robert-Nicoud instrumented for the share of developed land with two variables reflecting natural amenities—average January temperature and the presence of a major coastline—and population density in 1880, which they interpret as reflecting the innate desirability of an area at a time long before regulations were in place.²² These instruments turn out to be strong predictors of the share of developed land in 1992, and using them, the predicted share of developed land has a positive and significant effect on regulation. Thus, they interpreted the positive correlation

²² More specifically, their argument is that natural amenities are not directly correlated with supply restrictions because the types of regulations that they examine do not include efforts to protect coasts and because warm temperatures are unlikely to be significantly related to the type of environmental considerations that typically induce planning. Saiz (2010) did, in fact, document a strong relationship between geographic constraints, which tend to be in coastal areas, and regulation. He interpreted this relationship as causal. However, this result does not undermine the validity of amenities as instruments for the share of developable land as long as the reason why coastal areas have tighter regulation is because they have less developable land.

between regulation and developed land as reflecting the stronger demand to regulate in high-amenity places.²³

As discussed above, [Saiz \(2010\)](#) followed a different empirical strategy using geographic features of the landscape to measure the supply of buildable land and hence the restrictiveness of housing supply regulation. [Saiz \(2010\)](#) found that metropolitan areas with a larger fraction of land lost to geographic constraints tend to have stricter housing supply regulations as measured by the WRLURI. Citing the models discussed above, he then argued that homeowners have a stronger incentive to protect their housing investment (through regulation) where land values are high initially and that geographic constraints lead to important initial differences in land values. In keeping with this interpretation, [Saiz \(2010\)](#) found that the relationship between geographic constraints and regulatory constraints does not hold in areas suffering from weak long-term demand (defined as the bottom quartile of urban growth between 1940 and 1970). His proposed explanation is that geographic constraints are not likely to bind in low-demand places and, therefore, they are not likely to affect the incentive to regulate.

In contrast to the results documented by [Saiz \(2010\)](#) and [Hilber and Robert-Nicoud \(2013\)](#), a few other papers have found the opposite correlation of historical density patterns with regulation. In their analysis of townships in the Greater Boston area, [Glaeser and Ward \(2009\)](#) showed that higher housing density in 1940 and 1915 is associated with less restrictive residential zoning in the late 1990s/early 2000s, as measured by minimum lot size requirements. This relationship is quite strong in that density in 1940 can explain 68% of the variation in current minimum lot sizes. Going back even further, these authors showed that towns with larger minimum lot sizes had a larger fraction of forested area in 1885. Their interpretation is that historical patterns of deforestation reflect agricultural productivity. Locations with more productive land were developed first and had higher land values, resulting in higher density for many years to come and eventually to regulations that allowed for higher residential density.

[Evenson and Wheaton \(2003\)](#) also used data from Massachusetts to explore the connection between land use regulation and past development patterns. They measured actual land use in 1999 from aerial photographs taken of each town by the Massachusetts Office of Geographic and Environmental Information and obtained parcel-level information on zoning bylaws and the amount of land protected from development from the same institution. The data on open land are combined with minimum lot sizes and floor-to-area ratios to determine the maximum possible number of new buildings

²³ The economic magnitude of this effect is a matter of opinion. It seems large in that assigning San Francisco's attributes to Salt Lake City would move the latter from the 56th most regulated area (out of 93) to the 35th most regulated. On the other hand, a one standard deviation change in the share of developed land is associated with only one-third standard deviation stricter regulation, which is about the same magnitude that the authors found for homeownership, as described above.

in each town under current zoning regulation. They found that towns with relatively high housing density in 1999 tend to zone for higher future housing densities. In addition, towns with a larger fraction of developed land devoted to residential use tend to zone a larger fraction of open land for residential use. Like Glaeser and Ward (2009), they found that zoning regulations follow past land use patterns. They also showed that for all but a handful of towns very close to Boston, future residential building is restricted to a lower density than the current density (i.e., the slope of the regression line of zoned density on current density is positive but less than one). This result fits Fischel's ideas that residents of suburban jurisdictions are more likely to focus on the negative aspects of allowing new construction, while residents of urban jurisdictions where many businesses are located are more sensitive to the potential benefits of additional development.

The contrasting findings of Hilber and Robert-Nicoud and Saiz on the one hand with those of Glaeser and Ward and Evenson and Wheaton on the other suggest that much more work is needed to understand the connection between density patterns and regulation. One difference between these two sets of results is that they use very different data. The first two studies measure regulation using the WRLURI, which covers many aspects of the regulatory process including the actors in the development process, statutory limits on residential construction, and delays in permit approval, whereas the latter two studies focus on minimum lot size zoning.²⁴ Another important difference between these two sets of studies is that the first examines patterns across metropolitan areas, while the second examines patterns across jurisdictions within a single state or metropolitan area. As we discuss below, the incentives for regulating new development could be different in these two settings. In addition to further empirical analysis examining the relationship between the supply of buildable land and regulation, it would be helpful to consider other models that might help explain why zoning patterns seem to be correlated with past development.²⁵

19.3.4 Regulation in a multicomunity setting: Sorting and strategic interactions

Thus far, we have mostly focused on land use decisions in the context of a single community in isolation.²⁶ But what happens when the population moves freely across a set of local jurisdictions, each with the ability to set its own land use policy? A number of papers have investigated the use of zoning to induce sorting of households across jurisdictions.

²⁴ It could be that communities set lot sizes in a way that allows new construction to follow the density of existing buildings but restrict new development in other ways to compensate.

²⁵ One step in this direction is Evenson and Wheaton (2002, unpublished working paper), who predict an inverse correlation between existing residential density and restrictions on future development. The intuition is that for a given population size, higher existing density implies lower aggregate consumption of land. When existing residents consume less land, the capital gains from open-space creation become smaller relative to the cost of purchasing that space, so they will set aside less open land for public use.

²⁶ The primary exception is the discussion of Hilber and Robert-Nicoud (2013), whose model incorporates mobility across jurisdictions.

One such motive that has been discussed frequently in the literature is the “fiscal” motive, which refers to a desire to limit the fiscal externality that arises when households pay tax proportional to their income but share the benefits of public expenditures equally (Hamilton, 1978). In that case, zoning restrictions such as minimum lot sizes allow households to sort themselves into communities with other households that have similar preferences for the provision of a public good. Epple et al. (1988) developed a model where uncertainty about the type of household that will arrive in a metropolitan area can cause homeowners in a community to enact zoning in order to maximize the net tax contribution of new entrants. Similarly, Calabrese et al. (2007) showed how homeowners will choose to impose a lower bound on housing consumption to induce sorting across communities and consequently more efficient provision of a local public good.

Rolleston (1987) argued that fiscal considerations will cause communities to compete with one another to maintain or improve their fiscal advantage relative to other communities in their region. Therefore, zoning should be more restrictive in an area with a “healthier” fiscal situation relative to others. In a sample of 360 communities in northeastern New Jersey, she created a measure of the “fiscal capacity” of a community based on per capita income; per capita market value of residential taxable property, farm land, and vacant land; and per capita market value of commercial, industrial, and telephone/telegraph property. She found that relative fiscal capacity is positively related to the restrictiveness of residential zoning, as measured by an index of minimum lot sizes for vacant developable residential land.

Bates and Santerre (1994) also found evidence in favor of fiscal motives for zoning in a sample of 132 municipalities in Connecticut. Specifically, they found that communities are likely to zone a lower fraction of land for residential purposes when they have a large nonresidential tax base, which they argued reflects a larger opportunity cost of residential development. They also found that the fraction of vacant land zoned for residential use is negatively related to past population growth, which they interpreted as a fiscal motivation because rapid population growth can “impose increasing public service costs on entrenched residents.” On the other hand, they found little correlation between these two fiscal-motivation variables and minimum lot sizes. They also found no relationship between the total market value of taxable property, another variable reflecting the fiscal motive for zoning, and either measure of zoning restrictiveness.

One problem with interpreting the results found by Rolleston (1987) and Bates and Santerre (1994) is that the variables used to measure the fiscal capacity of a jurisdiction could also reflect a number of other factors that influence regulation. Moreover, these measures might be affected by the amount of regulation in the area. Lutz (forthcoming) mitigates these identification issues by examining the effect of a statewide policy that had a differential impact on the fiscal situation of local governments. Specifically, in an attempt to equalize funding for education, a reform in New Hampshire began granting payments to towns on the basis of their property wealth. He shows that towns with a larger grant were more likely to impose growth management regulations,

consistent with the idea that a stronger fiscal situation will cause a community to use regulation to restrict new entry.²⁷

Another motivation for zoning that is frequently discussed in the literature is that households may use zoning to prevent minorities and/or poor households from moving in for reasons other than fiscal motives. This often is referred to as an “exclusionary” motive for zoning in the literature.²⁸ Because minorities tend to have lower incomes, minimum lot size requirements or regulations imposing a floor on the quality of homes can restrict access to potential minority residents and potential low-income residents. In a sample of tracts in Santa Clara County, CA, [Pogodzinski and Sass \(1994\)](#) found that minimum lot sizes are larger in tracts with higher median incomes, which could be due to either a fiscal or an exclusionary motive. On the other hand, it could also be due to the fact that rich households are more adept at influencing land use policy.

[Rolleston \(1987\)](#) looked for evidence of exclusionary motives by comparing community characteristics to the characteristics of nearby communities. She found that communities with a lower percentage minority relative to other communities tend to have more restrictive residential zoning, consistent with the notion that residents are trying to keep out other races. [Bates and Santerre \(1994\)](#) found similar evidence for exclusionary zoning in that a smaller fraction of land is zoned for residential use when poverty in the nearest central city is higher relative to poverty in the zoned community. Also, [Shertzer et al. \(2014\)](#) document that Chicago neighborhoods with a larger share of black residents in 1922 were more likely to be zoned for manufacturing use and denser structures in a 1923 comprehensive zoning ordinance. In his survey of the literature on fiscal and exclusionary zoning, [Ihlanfeldt \(2004\)](#) concluded that most studies find evidence supportive of the fiscal motive to restrict land use, while evidence on exclusionary motives is mixed. But he pointed out that it is difficult to distinguish between sorting by income, which could be interpreted as an outcome of fiscal considerations, and sorting by race, which is more likely exclusionary. He called for more empirical studies to distinguish more clearly between income and race, something that is undoubtedly needed. Moreover, further research should attempt to distinguish sorting that arises from the fiscal and exclusionary motives from other factors that might cause households to sort into communities.

Once a model of regulation is expanded to consider population flows across jurisdictions, the land use policies in one community will affect the decisions in surrounding communities. [Brueckner \(1995\)](#) illustrated this idea in a model where local planners seek to maximize social welfare, which includes the utility of residents who pay rent and of absentee landlords who collect rent. Free mobility of residents across communities means

²⁷ Towns receiving large grants also experienced a surge in building activity, so the growth management results are also consistent with the possibility that population inflows lead to an increase in regulation.

²⁸ [Lynch and Rasmussen \(2004\)](#) and [Cervero and Duncan \(2004\)](#) found that property values are negatively associated with racial diversity in the neighborhood, suggesting that exclusionary motives for zoning could stem from the desire to preserve one's property value.

that planners must account for how a growth control will affect the relative demand to live in their community. He also assumed that city size is nonnegligible relative to the urban system, so that population diverted by a restriction in one city will increase demand pressure in other cities. In the most general form of his model, the effect of one city's land use policy on another's is ambiguous. But assuming that individuals have Leontief preferences over housing and all other goods, he showed that growth controls in a city will be strategic complements for marginal movements away from a symmetrical equilibrium. That is, regulation in one city will become stricter (looser) if it becomes stricter (looser) in a neighboring city.²⁹

Helsley and Strange (1995) also modeled strategic interactions among local communities. In their work, growth controls raise property values not only because they restrict the housing supply but also because residents' utility is decreasing in the population size of their community. Consequently, communities have an incentive to exclude potential residents because this reduces utility in other communities and make residents more willing to pay to live in the restricted community. Because regulations lead to higher rents, even developers may support the adoption of strong growth controls. This result is intriguing because it suggests a reason for why the relationship between homeownership and regulation has been hard to find empirically—namely, that developers have incentives to support regulation in the same types of communities that homeowners do. The model of Hilber and Robert-Nicoud (2013) discussed above also allows for household mobility across a set of jurisdictions, with the result that regulation in one jurisdiction is weaker when amenities are more desirable (and thus housing supply regulation is stronger) in other communities.

Turning to empirical evidence on these interjurisdictional strategic interactions, Brueckner (1998) showed that cities in California are more likely to impose stringent growth controls when a nearby city has done so. Similarly, Lenon et al. (1996) found that minimum lot sizes in Connecticut townships are larger when lot sizes in neighboring townships are larger. Because these studies are all based on cross-sectional correlations, it is possible that their results could reflect similarities in the population or economic environment of jurisdictions that are in close geographic proximity. This is yet another example of the benefits that will arise when panel data are available, as one could potentially observe how the adoption of regulations in a community results in changes in regulations in other communities.

To summarize the literature on the determinants of housing supply regulation, it seems that zoning does not simply mirror the land use that would result in an unconstrained market, but reflects the intention of local governments to alter land use and control the amount and quality of residential structures in their area. Theories delving into how local politics shape regulation have focused on the incentives for homeowners to

²⁹ Brueckner (1995) also showed that this strategic complementarity reverses for very large movements away from the initial equilibrium.

restrict development. Geographic constraints also appear to influence regulation by altering the price of land and thus changing the incentives to restrict development. In addition, because regulation can lead to sorting of households across communities, research has examined motives for zoning that relate to preventing certain types of households from entering a community. Much of the empirical analysis evaluating all of these theories is based on cross-sectional relationships, raising concern about omitted variables and reverse causality. The use of panel data on regulation and/or instruments would allow for a cleaner assessment of all of these theories.

19.4. EFFECTS OF REGULATION

19.4.1 Effects on the price and quantity of housing

We now turn from the determinants of housing supply regulation to its effects. Much of the research concerning these effects has focused on the housing market. In simple models with only one city, regulation leads to higher rents and house prices and less construction because it reduces the elasticity of housing supply. In particular, regulation increases the marginal cost of construction, both directly through the fees and time costs and indirectly by requiring construction to follow certain forms (lot size and setback requirements, for example) that the builder would not otherwise choose and by creating uncertainty about project approval. Also, some types of regulation such as growth controls effectively make the marginal cost of housing infinite by constraining the total number of housing units allowed. Brueckner (2009) presented a series of simple models to illustrate the predicted effects of various types of regulation, including urban growth boundaries, floor area ratio restrictions, and cost-increasing measures such as delays or uncertainty, on the price and quantity of housing. In most cases, regulation leads to higher land rents, higher house prices, and a smaller housing stock.

The predicted effects of regulation become less obvious in the context of a system of cities where households move freely among them. In this case, utility must be equal across cities so population flows can erode away price differences across locations. In the model developed by Helsley and Strange (1995), equilibrium land rent in a community will be a function of land rent in a “passive” community that does not set growth controls, the population of the passive community, and the disutility caused by population congestion. In the case where local amenities are not eroded by population, land rents are equal in all communities. Land rents are still higher in all communities compared with the case of no regulation because growth controls increase the population in the passive community, thereby raising rents through greater demand for housing.³⁰ In the case where congestion

³⁰ Pollakowski and Wachter (1990) showed that density restrictions in one planning area of Montgomery County, MD, raise house prices in adjacent planning areas, illustrating the idea that mobility arbitrages away price differentials.

does reduce utility, rents are higher in communities with growth controls because households value living in communities with a lower population. Thus, there are price differences across communities, but they reflect the amenity value of growth controls, not the lower elasticity of housing supply. Just as in the single-city case, regulation will reduce the number of housing units in controlled cities.

Numerous papers have examined the empirical relationship between regulation, house prices, and construction. Many of these have used cross-sectional data on communities within a single metropolitan area so they might underestimate the aggregate effect on house prices.³¹ Others have compared average degrees of regulation across metropolitan areas, in which case population flows are less likely to erode away price differentials.³² Although this logic suggests that the estimated effects of regulation on house prices should be larger in cross-metropolitan analyses, metropolitan-level measures of regulation are constructed from averages over only some jurisdictions within the metropolitan area so estimates at the metropolitan level might suffer from more downward bias due to measurement error than those based on jurisdiction-level data where regulations are measured more accurately.

The vast majority of studies have found that locations with more regulation have higher house prices and less construction. The magnitudes of these estimated effects are difficult to compare across studies because each one uses different measures of regulation—a few examples include indicators for the presence of a particular regulation, the number of types of regulation used in an area, the fraction of vacant land restricted from development, the number of months needed to complete the permitting process, or indexes of various regulations. Nevertheless, it is notable that many estimates seem to be quite large. As noted earlier, [Katz and Rosen \(1987\)](#) found that in the San Francisco Bay Area, house prices are at least 17% higher in jurisdictions with at least one formal growth management program. In [Malpezzi's \(1996\)](#) study of 56 metropolitan areas, increasing the level of regulation from its average by one standard deviation is associated with 22% higher house prices and 11% lower construction. Also in a cross-section of metropolitan areas, [Mayer and Somerville \(2000\)](#) found that increasing the time to receive a subdivision approval by only 1 month reduces building permits by 10%. In their survey of the literature, [Quigley and Rosenthal \(2005\)](#) provided a list of 40 articles estimating the effect of regulation on the price and quantity of housing and their results.

One important issue when using a cross-section of locations to infer the effect of regulation on housing markets is that regulations are not distributed randomly across

³¹ Examples include [Katz and Rosen \(1987\)](#), [Pollakowski and Wachter \(1990\)](#), and [Quigley and Raphael \(2005\)](#).

³² Examples include [Malpezzi \(1996\)](#), [Mayer and Somerville \(2000\)](#), [Segal and Srinivasan \(1985\)](#), and [Black and Hohen \(1985\)](#).

locations. The theories discussed in [Section 19.3](#) suggest that higher house prices could induce stricter regulation by giving homeowners a greater incentive to protect their property values. In this case, high house prices lead to stricter regulation, rather than vice versa. Also, faster growing areas could implement growth restrictions to reduce congestion or preserve local amenities, making the negative effect of regulation on construction more difficult to observe in the data. Because regulations change infrequently and tend to reflect historical land use patterns, it is not obvious that simply using lagged values of regulation can alleviate concerns about reverse causality. Another difficulty with identifying the effects of regulation is that many unobservable variables, such as the amenity value of a location, could be correlated with housing supply regulation and having independent effects on house prices and quantities.

As we mentioned above, a few studies have constructed panel datasets in order to mitigate these endogeneity problems. [Glaeser and Ward \(2009\)](#) used data on the adoption of regulations that affect new construction—specifically the institution of wetlands bylaws, septic rules, and rules regarding roads and sidewalks in subdivisions—in Massachusetts townships. Using data on residential building permits from 1980 to 2002, in specifications without town fixed effects, they found that wetland bylaws and subdivision rules reduce new construction by about 6% and 15%, respectively. These estimates increase to about 10% and 22% when town fixed effects are included, suggesting that these regulations tended to be implemented in areas with more-than-average construction activity (as hypothesized above). However, the standard errors of these estimates are fairly large—not only are they likely to be insignificantly different from one another, but only the estimated effect of subdivision rules in the fixed effects specification is significantly different from zero.³³

Glaeser and Ward did not present fixed effects specifications with house prices as the outcome of interest, perhaps because they only have house price data for years after most regulations were adopted. Because the regulations in their sample are uncorrelated with most town characteristics except historical density, they argued that these regulations can be treated as random conditional on density so they present cross-sectional regressions of house prices on regulation. Controlling for only a small subset of town characteristics, they found a small positive correlation between an index of regulation (made up of wetlands bylaws, septic rules, and subdivision rules) and house prices. But this correlation disappears when controlling for additional town characteristics. These results are consistent with the [Helsley and Strange \(1995\)](#) model, in that price differences are negligible between towns that are close substitutes of one another.

³³ In a fixed effects specification, they also found that cluster zoning appears to *increase* new construction, which they interpreted as suggesting that these rules can encourage construction by directing it to certain areas, but again, this estimate is not significantly different from zero.

Zabel and Dalton (2011) also used time series data from Massachusetts to examine the effects of housing supply regulation. In their case, they do not observe changes in regulation directly. However, they inferred changes in minimum lot size restrictions by examining changes in the 25th percentile of observed lot sizes of new construction. They argued that the median lot size might not reflect minimum lot size restrictions because these restrictions may only affect the bottom of the lot size distribution but that the 25th percentile likely will reflect these regulations. After assembling data on the 25th percentile of new construction lot sizes in each town, they used a structural break framework to identify sharp changes in these series that likely reflect regulatory changes. In the cross-section, they found that minimum lot sizes are uncorrelated with house prices. In specifications with zoning district fixed effects, they found that an increase in minimum lot size restrictions of 1 acre (1.5 standard deviations) raises house prices by about 9%.³⁴

A third paper that uses panel data on regulation is Jackson (2014), who constructed a panel of cities in California from 1970 to 1995 based on a survey carried out by Glickfeld and Levine (1992), which included information on when regulations were first implemented. He found that an additional regulation reduces residential permits by 4–8%, a smaller effect than what others have estimated in a cross-section with the same data.

Rather than using panel data, Ihlanfeldt (2007) instrumented for land use restrictions with past community characteristics that he argued are correlated with a homeowner's ability to influence the local regulatory process. Specifically, his instruments included educational attainment, the fraction of residents older than 55, race, homeownership, and average income. He showed that these variables are strong predictors of regulation, as measured by the number of restrictive land management techniques used in a jurisdiction. Although one might be concerned that these community characteristics might be correlated with other factors that influence house prices independently of regulation, he showed that this set of instruments passes the standard overidentification tests. It is curious that these instruments work well in his sample of housing units in Florida, since Glaeser and Ward (2009) found that they do not work for townships in Massachusetts. In a sample of sales transactions in Florida, he found that predicted regulation raises the price of single-family homes substantially.³⁵ He also found that this effect is smaller in counties containing more cities, consistent with the idea that in areas where local residents have

³⁴ They also provided a nice illustration of the idea that migration across communities can erode away price differentials. In particular, for towns that are estimated to have little zoning power because they have similar amenities to others (as calculated from fixed effects regressions of house prices on structural characteristics, year fixed effects, and town fixed effects), minimum lot sizes only boost house prices by a small amount. But when a community has greater zoning power, the effect of a 1 acre increase in minimum lot size on house prices is in the range of 15–20%.

³⁵ He found a smaller effect of regulation on house prices when regulation is assumed to be exogenous, which is somewhat puzzling since most theories would suggest that regulation should be stronger in high-priced areas, biasing the estimated effect using OLS upward.

more choices over where to reside, population flows will erode differences in house prices across locations.³⁶

So far, our discussion has focused on regulations that are designed to restrict the housing supply or otherwise shape residential development. Impact fees do not quite fit this description in that they are fees charged to developers under the justification of being used to finance the public infrastructure that will be required by the new development. Compared with other ways of financing infrastructure, such fees will reduce the amount of residential development, but the predicted effects on land values are ambiguous (Brueckner, 1997). McFarlane (1999) pointed out that the way the fee is structured, for example, whether it is on land, on housing, or on the value of the developed land, matters for these predicted effects. On the empirical side, Skidmore and Peddle (1998) found that residential development in DuPage County, IL, fell substantially after the imposition of an impact fee. By contrast, Burge and Ihlanfeldt (2006a,b) developed a model to show that impact fees can boost construction by substituting for exclusionary regulations and decreasing uncertainty about project approval. In a panel of counties in Florida, they document a positive relationship between construction of most types of suburban single-family homes and impact fees other than for water and sewer system improvements. These same fees expand multifamily construction in inner suburban areas, but not in central cities or in the outer suburbs.

Although most empirical work has focused on the effects of regulation on the levels of house prices and quantities, a few papers have examined the effect on the elasticity of housing supply by looking at housing market responses to shocks to housing demand. Saks (2008) proposed that shocks to local labor demand should shift the demand for housing as well as the demand for labor. Following Bartik (1991), she calculated shocks to labor demand in a metropolitan area using its industrial composition and national changes in employment by industry. She found that increases in housing demand result in a larger increase in house prices and less construction in metropolitan areas with tighter housing supply restrictions, as measured by an index of regulatory variables created from a number of sources. However, to the extent that these regulations might be correlated with geographic constraints as in Saiz (2010), her results might reflect the effects of natural constraints on the elasticity of housing supply as well as regulatory barriers.

³⁶ In another work, Chakraborty et al. (2010) instrumented for the density of zoned housing units—specifically, the number of units allowed in zones that permit 8 units per acre or more—using the percentage of the population that is white in 1960 and distance to the central city in a sample of suburban jurisdictions in six metropolitan areas. Other potential instruments, including lagged values of median household income, educational attainment, government revenue or expenditure per capita, and government fractionalization, had no relationship with zoned density. They found a small positive relationship between higher allowed densities and the change in multifamily construction, suggesting that density restrictions impede the development of multifamily housing. However, their two instruments have *t*-statistics between 2 and 5, and there are no test results reported for weak instruments or for exogeneity.

In a sample covering 353 local planning authorities (LPAs) in England, [Hilber and Vermeulen \(2013\)](#) found that a similar labor demand shock leads to larger increases in house prices in more regulated areas. Their analysis instruments for regulation with historic population density, the share of votes for the Labor Party in 1983, and a national policy reform that caused some LPAs to change their permit approval process more than others. It also controls for geographic constraints, as measured by the difference between the highest and lowest altitude in the LPA. In that sense, it provides a cleaner measure of the effect of regulation than [Saks \(2008\)](#). They showed that the estimated effect of regulation on the elasticity of housing supply is smaller when treating regulation as exogenous, suggesting that in their sample, regulation tends to be stricter in areas that have a less elastic housing supply for other reasons.

Instead of using instruments to estimate the elasticity of housing supply, a few papers have used time series models. [Wheaton et al. \(2014\)](#) estimate the elasticity of housing supply for 68 metropolitan areas using a vector error-correction model of house prices and the housing stock. They find that the long-run elasticity of housing supply is negatively correlated with the WRLURI. Also, in a vector error-correction framework, [Harter-Dreiman \(2004\)](#) found that changes in local income are correlated with larger house price increases in constrained relative to unconstrained metropolitan areas, where constraints are measured using the [Linneman et al. \(1990\)](#) survey from the late 1980s as published in [Malpezzi \(1996\)](#). And [Green et al. \(2005\)](#) found that the time series correlation between construction and changes in house prices, which they interpreted as an elasticity of housing supply, is lower in metropolitan areas where regulation is high, as measured by the same survey from [Linneman et al. \(1990\)](#).

If regulation reduces the elasticity of housing supply, then one would expect house prices to be more volatile and construction to be less volatile, all else equal. Indeed, [Malpezzi and Wachter \(2005\)](#) showed that metropolitan areas with more regulation (as measured by seven variables from the [Linneman et al., 1990](#) survey of the late 1980s) had a higher standard deviation of house price changes as measured by repeat-sale house price indexes from 1979 to 1996. [Glaeser et al. \(2008\)](#) found that house price increases were larger during the 1980s and smaller during the early 1990s in metropolitan areas with tighter geographic constraints on the housing supply, suggesting that house price bubbles were more pronounced in areas with an inelastic housing supply. Although they focused on geographic constraints, those authors noted that these geographic constraints are highly correlated with regulation.

One concern with these studies is that the greater volatility of house prices in more regulated areas is the result of a greater volatility of shocks, rather than due to a steeper supply curve. Malpezzi and Wachter controlled for the standard deviations of annual changes in employment and annual changes in real income per capita, but these two variables may not pick up the volatility of other shocks. [Paciorek \(2013\)](#) addressed this concern by estimating and simulating a dynamic structural model. In his model, changes in

house prices relative to construction costs are a function of investment (i.e., construction) and density. The interaction of investment and density with the degree of regulation shows how supply constraints alter the elasticity of house prices with respect to each of these variables. He estimated the parameters of the model using two shocks to housing demand to identify the elasticity of housing supply—a Bartik-style predicted employment shock and a predicted migration shock calculated using annual outmigration from metropolitan areas and average migration flows across pairs of metropolitan areas. Based on these estimated elasticities, he then simulated the evolution of prices and construction in response to 100 randomly drawn demand shocks. He found that house price responses to a demand shock are more volatile in areas with more regulation. Conversely, construction is less volatile in these areas. These differences are sizeable, as going from the 90th percentile of regulation to the 10th percentile of regulation reduces volatility of prices by 20% and increases the volatility of investment by 30%.

One paper that stands at odds with those cited above is [Davidoff \(2013\)](#), who pointed out that in the most recent house price cycle, metropolitan areas with relatively little regulatory and geographic barriers, such as Las Vegas and Miami, experienced the most pronounced booms and busts. Notably, most of the metropolitan areas with large house price bubbles, defined as the difference between house price growth in 2000–2007 and 2007–2010, experienced very large amounts of construction and small house price increases in the 1980s. Since regulation evolves slowly over time, it does not seem plausible that big changes in regulation can account for the differential behavior over these two cycles. Using state fixed effects to control for differences in demand shocks across locations, he further showed that regulation and geographic constraints are unrelated to the magnitude of the recent house price bubble. These results suggest that it was considerable fluctuations in housing demand across locations, rather than a lower elasticity of housing supply, that can explain variation in housing market fluctuations in the most recent cycle. In another chapter in this volume, Glaeser discussed housing bubbles in much greater detail.

In summary, most models predict that regulation should reduce the elasticity of housing supply, resulting in a smaller stock of housing, higher house prices, greater volatility of house prices, and less volatility of new construction. Although many empirical studies do find evidence of these relationships, they are mainly cross-sectional studies and consequently are potentially biased by omitted variables and reverse causality. A few studies have used panel data or community characteristics in order to mitigate these issues, but much more should be done along these lines.

19.4.2 Effects on urban form and homeownership

Besides its effects on the quantity of housing, housing supply regulation can influence other aspects of urban form. [Mills \(2005\)](#) developed a model to show that density controls

contribute to urban sprawl by forcing development to extend further than it would in a competitive equilibrium. [Bertaud and Brueckner \(2005\)](#) demonstrated a similar result for building height restrictions.

Of course, some types of regulation are meant to prevent sprawl by prohibiting development outside a particular boundary. Portland is one frequently used example. [Quigley and Swoboda \(2007\)](#) described a model where an urban growth boundary raises the density of development within the permitted area while reducing the size of the urban area relative to the unregulated case. Other types of regulations, such as maximum lot sizes or minimum density restrictions, limits on the number of building permits, or impact fees, can also reduce sprawl ([Pasha, 1992](#); [Song and Zenou, 2006](#); [Geshkov and DeSalvo, 2012](#)).

In a sample of 182 urbanized areas in 2000, [Geshkov and DeSalvo \(2012\)](#) found that urban areas are larger when county regulations include minimum lot sizes and maximum floor area ratios (i.e., height restriction), while urban areas are smaller when regulations include maximum lot size, a maximum number of building permits, minimum density, or impact fees. They found little correlation between the presence of an urban growth boundary and the size of the urban area, which they speculated could be either because these boundaries are not binding or because they do not completely surround the urban area, allowing development to push out in some directions. Another issue is that urban growth boundaries are likely to be used in areas with fast population growth where residents are worried about preserving local amenities, so places that use urban boundaries may be larger on average than areas that do not.

Another aspect of urban form than can be influenced by land use policy is the fraction of single-family versus multifamily units. As above, not all types of regulation will have the same effect. Minimum lot sizes or height restrictions are likely to reduce the share of multifamily housing, while urban growth boundaries might increase the share of multifamily housing by concentrating development in a smaller geographic area ([Pendall, 2000](#); [Nelson et al., 2004](#)). [Hilber et al. \(2014\)](#) develop a formal model to show how regulations can affect the supply of different types of housing, even though all households prefer single-family floor space to multifamily floor space. In their model, as long as the amount of land available for redevelopment in the central city is large enough, an unexpected increase in household income will raise the amount of multifamily construction by more than single-family construction, because it will lead to more construction in areas where land is relatively expensive. However, in the presence of regulations such as minimum lot sizes or restrictions against multifamily housing, the construction response to an income shock could be shifted towards areas where it is more profitable to construct single-family housing.

Several studies have documented a positive correlation between regulations that restrict housing unit density and the share of single-family versus multifamily construction ([Shlay and Rossi, 1981](#); [Pendall, 2000](#)). [Quigley and Raphael \(2005\)](#) found a positive correlation between the number of land use regulations in a municipality—including

regulations that restrict the density and height of nonresidential buildings, regulations that require “adequate” preexisting service levels for residential and commercial development, and regulations that require voter approval or supermajority city council votes in order to increase density—and new single-family construction, while they found no correlation between these same regulations and new multifamily construction. Hilber et al. (2014) assess the effect of regulation on the types of housing demanded using income shocks similar to the Bartik-style employment shares to proxy for housing demand. They find that income shocks are typically associated with an increase in the share of newly constructed multifamily housing, especially in the central city of the metropolitan area. By contrast, in metropolitan areas where housing supply regulation is stronger than average, there is no relationship between income shocks and the type of newly constructed housing. They concluded that regulation does indeed alter the type of housing unit constructed.

Another potential housing-related effect of regulation is on homeownership. There are several channels through which this might occur. The first is through the relative cost of owning versus renting. In models where house prices are the discounted sum of expected future rents, as in Ortalo-Magne and Prat (2014), restrictions on the supply of housing will raise prices in proportion to the increase in rents, leaving homeownership unchanged. But regulation could change other components of the user cost of housing. For example, Fischel’s homevoter hypothesis suggests regulation could reduce the risk of a drop in home values associated with undesirable new construction, and a decrease in the risk premium associated with investing in housing would boost homeownership if homeowners are more risk-averse than landlords. Moreover, to the extent that managing single-family rental property is more costly than managing multifamily rental property, one might expect regulations that restrict the supply of multifamily housing to raise the homeownership rate.³⁷

Only a few papers have examined the effect of regulation on homeownership empirically. As we have discussed at length above, one should expect homeownership to lead to tighter regulation, so any empirical analysis must confront the possibility of reverse causality. In their sample of Census tracts in Chicago, Shlay and Rossi (1981) found that tracts with more regulation experienced larger increases in the number of owner-occupied housing units and larger decreases in the number of renter-occupied units between 1960 and 1970, suggesting an increase in homeownership. Their use of changes in the housing stock ameliorates concerns about reverse causality, but one still could worry that regulation reflects past construction patterns, which could in turn reflect

³⁷ Although Ortalo-Magne and Prat did not examine the effect of regulation on investment in owner-occupied housing, a decline in the risk of a decrease in house prices could be modeled as a decline in the variance of shocks to rent, while an increase in the cost of managing rental housing could be modeled as an increase in the homeownership premium.

demand for owner-occupied housing. In a cross-section of metropolitan areas, [Malpezzi \(1996\)](#) found little correlation of regulation with the homeownership rate. However, he also found that regulation raises house prices by more than it raises rents. Based on the correlations of prices and rents with the homeownership rate, he concluded that regulation reduces homeownership.

In summary, the predicted effects of regulation on urban form depend somewhat on the type of regulation being considered. It seems that minimum lot sizes and height restrictions contribute to sprawl, while maximum lot sizes, limits on construction, minimum density requirements, and impact fees reduce sprawl ([Geshkov and DeSalvo, 2012](#)). Meanwhile, density restrictions also increase the share of single-family housing ([Shlay and Rossi, 1981](#); [Pendall, 2000](#)). But as we have now said repeatedly, this research is based on cross-sectional evidence and we could learn more from panel data and other ways to control for omitted variables that might be correlated with both regulation and urban form. Turning to homeownership, we are aware of little research on the effect of regulation on homeownership—either theoretical or empirical—and this topic also would be useful to explore in future work.

19.4.3 Effects beyond housing markets

By altering the distribution of house prices across locations, housing supply regulation can influence the location choices of different types of households and hence affect the typical characteristics of households residing in any given community. Perhaps, the most obvious household characteristic to consider is income, because richer households will be more able to bear the higher housing costs caused by regulation ([Downs, 1973](#); [Ellickson, 1977](#); [Neiman, 1980](#); [Gyourko and Voith, 2002](#); [Gyourko et al., 2013](#)). Of course, empirically, it is difficult to identify the effect of regulation on household income separately from the increased likelihood that higher-income households vote to enact stricter regulation.

One way to get at the effect of regulation on income sorting is to look at changes in neighborhood income after a regulation is enacted. For example, [Kahn et al. \(2010\)](#) showed that after the creation of a coastal boundary zone to regulate construction near the California coastline, household income rose faster in Census tracts inside the zone than outside the zone, even after controlling for distance to the coastline and a number of other tract characteristics. [Levine \(1999\)](#) found that, controlling for 1980 income, cities that enacted more growth control measures between 1979 and 1988 had higher incomes in 1990.³⁸

³⁸ [Shlay and Rossi \(1981\)](#) found that conditional on the 1960 income, suburban tracts in Chicago with a lower fraction of residential area zoned for high-density development had higher incomes in 1970. The correlation between regulation and income is much smaller for tracts in the central city.

Gyourko et al. (2013) identified income sorting across metropolitan areas using a national demand shock that is calculated from changes in the aggregate number of high-income households. They showed that a positive national demand shock results in a larger share of high-income residents in “superstar” metropolitan areas, which they defined as areas where the sum of house price and housing unit growth over the previous 20 years is above the median across metropolitan areas. Because highly regulated metropolitan areas tend to have above average price growth, many of the “superstar” areas in their sample tend to be highly regulated. Thus, the aggregate shock results in more high-income households in highly regulated areas.

Because income is correlated with many other household characteristics, one might imagine that housing supply regulation affects the geographic distribution of other household characteristics. One such characteristic that has received much attention is race. As mentioned above, some have argued that racial segregation may be a factor motivating the adoption of zoning regulations (Rolleston, 1987; Nelson et al., 2004). Indeed, a number of studies have found that more regulated communities experience slower growth in minority populations (Levine, 1999; Pendall, 2000; Quigley et al., 2004) or a drop in the share of minority populations (Donovan and Neiman, 1995). These results have been used to argue for the presence of the exclusionary motive for zoning, but caution is in order regarding this interpretation. The fact that zoning results in sorting does not necessarily mean that sorting was the intended consequence of the regulation. On the other hand, cases where patterns of regulation are consistent with an exclusionary motive certainly should be examined in further detail. Whether or not the effect is intended, the correlation of regulation with subsequent increases in racial segregation is an outcome of interest to policymakers. Nelson et al. (2004) found that metropolitan areas with urban containment boundaries experienced *smaller* increases in racial segregation over the 1990s. They take this result as evidence that these types of regulation are less exclusionary than other types of land use policy because they frequently require the adoption of policies that promote affordable housing within the urban zone.

Ganong and Shoag (2013) developed a model to show that housing supply regulation will lead to sorting by skills as long as low-skilled workers spend a disproportionate share of income on land. Moreover, the model predicts that housing supply constraints will cause income differentials across locations to persist by limiting migration from low-wage to high-wage areas. Based on this model, they posit that an increase in regulation in some states has led to more skill sorting across states while reducing income convergence across states. To measure changes over time in regulation at the state level, they counted the number of state appellate court cases that mention “land use” in each state and year and then created an index based on the cumulative number of land use-related court cases since 1940. Consistent with their theory, they found that 20-year population growth rates are only negatively associated with initial income levels in states with low levels

of regulation. In addition, high initial income levels only lead to migration-related declines in human capital in states with low levels of regulation.

The effects of housing supply regulation can extend beyond the sorting of households across locations. Because most people work in the same metropolitan area where they live, housing supply regulation can affect the supply of labor and hence the functioning of local labor markets. In particular, a less elastic supply of housing may imply a less elastic supply of labor. In support of this conjecture, [Saks \(2008\)](#) showed that a Bartik-style labor demand shock results in lower employment growth and higher wages in metropolitan areas with tighter housing supply regulation. Similarly, [Glaeser et al. \(2006a\)](#) found that regulation reduced the response of the local population to a local productivity shock. One would expect that this result would show up as a diminished response of migration to labor demand shocks. However, [Zabel \(2012\)](#) found that migration flows into and out of metropolitan areas in response to a demand shock are *larger* in areas with a lower elasticity of housing supply or a higher average initial price level. These results are not directly comparable to Saks's because Zabel measured the elasticity of housing supply with [Saiz's \(2010\)](#) estimates, which are based on geographic constraints, rather than an observed measure of housing supply regulation. Nevertheless, because regulation is correlated with geographic constraints and with the level of house prices, it is puzzling that migration flows do not seem to respond more to demand shocks in more elastic areas.

In summary, regulation can lead to household sorting by income, race, and other characteristics. Whether intended or not, it is important for policymakers to consider these effects because the makeup of the community can affect many other outcomes including the demand for local public services and the size of the local tax base. Regulation can also affect local labor markets by constraining the response of the labor supply to changes in labor demand. Since the migration of workers is the primary means through which local labor markets respond to local changes in demand ([Blanchard and Katz, 1992](#)), the constraints imposed by regulation could have a meaningful influence on the economic health of local communities.

19.5. WELFARE IMPLICATIONS OF REGULATION

Thus far, we have said nothing about the welfare implications of regulation. This topic is not straightforward because any model-based assessment of welfare relies on assumptions about what aspects of regulation and the local environment enter agents' utility functions and how they are valued. In [Ortalo-Magne and Prat \(2014\)](#), regulation reduces welfare because its only function is to reduce the number of housing units; it does not create any local amenities that would compensate for its negative effects. City residents do benefit from higher house prices when they sell their house and move away, but at the expense of higher rents until then.

In [Helsley and Strange \(1995\)](#), population density enters with a negative sign in residents' utility functions, so regulations do convey some benefit. Even so, they found that regulation reduces welfare on net because the negative effect on house prices outweighs the positive effect of lower density. The reason is that migration from the regulated to the unregulated community raises house prices and population density in the unregulated community, reducing utility. Since free migration flows imply equality of utility across locations, it must be the case that residents in the regulated community are worse off as well.³⁹

But this result would not hold if the unregulated community was large enough to absorb the extra population without an increase in density or house prices. In models where the regulated community is small, the amenity created by population controls is fully capitalized in land rents, raising the utility of homeowners and landlords and reducing the utility of owners of undeveloped land, with the net effect on aggregate welfare being ambiguous ([Brueckner, 1990](#); [Engle et al., 1992](#)).

Empirical efforts to gauge the efficiency of land use controls include [Glaeser et al. \(2005\)](#), who asked whether there are sufficient counterbalancing spillovers to justify the very high "regulatory tax" that they estimate in their study of the Manhattan condominium market. In the context of Manhattan, it cannot credibly be argued that adding another person would spoil the bucolic nature of the area. However, there are clear external costs associated with building up. For example, views of existing residents are blocked, there is more congestion of the city infrastructure, and new residents could create a fiscal burden on current citizens. The authors estimate the value of a view from the price differential of apartments on different floors of the same building and congestion externalities from regressing rents on average income and population size in a sample of 193 cities. Both yield relatively small cost estimates compared to the magnitude of the regulatory tax. Furthermore, they argued that incoming Manhattan residents would likely represent a positive fiscal transfer to the city because high house prices imply that new residents would likely be richer than the average existing residents, and at the same time, these migrants are likely to have fewer children than the average current resident. Thus, [Glaeser et al. \(2005\)](#) concluded that land use controls make the housing supply inefficiently tight in Manhattan.

A more recent and ambitious estimate of the welfare consequences of supply constraints is provided by [Turner et al. \(2014\)](#). They bring much more data to bear on the question by combining property-level data on land sales from a major industry data provider (CoStar) with the WRLURI constraint measure and the 2006 National Land Cover database. Ultimately, they are able to create a national dataset, yet see fine geographic detail.

³⁹ In particular, house prices in the regulated community are even higher than in the unregulated community, more than offsetting the benefit of the smaller population.

Conceptually, the authors showed that the causal impact of regulation on land value is composed of three effects: (a) an “own-lot effect” that reflects the cost of regulatory constraints on how land is used, (b) an “external effect” that reflects the value of regulation for owners of nearby land, and (c) a “supply effect” that reflects the effect of regulation on the supply of developable land. Their empirical strategy to identify each impact relies on discontinuities at the border between municipalities. Essentially, they compare sales prices of vacant land across municipal borders where the restrictiveness of regulation changes. The key identifying assumption is that mean differences in unobserved traits of land on each side of the relevant border are not correlated with differences in regulation across the same border. [Turner et al. \(2014\)](#) argued that the exogenous nature of many borders and their ability to control for a wide variety of covariates help ensure that assumption is not violated.

The authors report a statistically and economically large and negative “own-lot” effect. Land values are about one-third lower if local regulation as reflected in the WRLURI measure is one standard deviation higher.⁴⁰ The authors cannot reject that the “external” component is zero, and the standard errors are small enough that they can conclude with high confidence that it could not be positive enough to counterbalance the “own-lot” impact. They also find a modest reduction in the supply of developable land if the local regulatory environment is one standard deviation higher. In total, a one standard deviation increase in the index of regulation reduces land prices by 36%, so [Turner et al. \(2014\)](#) also conclude that welfare would be improved by relaxing land use regulation.

The study by [Turner et al. \(2014\)](#) represents a major advance in our knowledge of and insight into the welfare consequences of local land use controls. However, their analysis is not the final word on the subject, as the authors clearly recognize. Their reliance on discontinuities at borders is a great strength in terms of econometric execution, but the estimated effects need not generalize to further away land parcels, and much development occurs away from borders of already built up suburbs. Moreover, their analysis could underestimate the “external” benefit of regulation because it is based on comparing land prices within a few hundred meters, and the benefits of, say, reduced traffic and more open space might accrue to a much larger area.

With respect to this latter point, [Cheshire and Sheppard \(2002\)](#) analyzed the welfare effects of regulation in a manner that allows them to include the more geographically

⁴⁰ The finding of a negative “own-lot” effect suggests that regulation prevents landowners from putting their land to its optimal use. However, not all constraints on the form of land development need make landowners worse off. [Libecap and Lueck \(2011\)](#) concluded that the rectangular system of land demarcation raises land values substantially relative to the metes and bounds system, even though it constrains the shape of parcels. They argued that even though the rectangular system is more constraining, it creates more secure and exchangeable property rights by reducing enforcement costs, trading costs, and coordination costs of infrastructure investment.

dispersed benefits. In a sample of homes for sale in Reading, England, they estimated a hedonic regression and expenditure share equation to determine the value that households place on limitations on industrial land use and the provision of open space, two key amenities generated by the local planning system in the area that they studied. Although they estimated fairly substantial gross benefits from these amenities, in a calibration exercise, they found that net welfare would be improved by relaxing land use regulation because the resulting reduction in house prices would outweigh the loss of amenities.⁴¹ In particular, a modest relaxation would lead to an average increase in utility of 2% of household income, while a significant relaxation would result in an average increase of roughly 4% of income.

In summary, most models and empirical estimates suggest that regulation reduces aggregate welfare, on net. And the estimated effects are often economically large. Yet more remains to be done, particularly on measuring the benefits that regulation may impart to local residents. In addition, further research into precisely who benefits and who loses from regulation, as well as the welfare consequences of different types of regulation, would be incredibly useful.

19.6. CONCLUSION

Despite the rapidly growing body of research on the regulation of the housing supply, much is still unknown about its causes and effects. One reason for our imperfect understanding is that most empirical work is based on cross-sectional data. Since regulation is likely correlated with many unobserved characteristics of the local geography, population, and economic environment, it is difficult to estimate causal effects from a cross-section alone. Throughout this chapter, we have pointed out examples where panel data would allow for much cleaner analysis of both the causes and effects of regulation. However, panel data would be extremely costly to collect because surveys must be very detailed in order to capture the wide variety of regulations that exist in different locations. In addition to more surveys, researchers should explore other readily available data that might shed light on the degree of regulation. It would also be fruitful to think of creative instruments for regulation and the potential determinants of regulation, which would allow for better causal estimation without the use of panel data.

Another unsatisfying aspect of current data on housing supply regulation is that the vast majority of data are from the United States. One notable exception is [Hilber and Vermeulen \(2013\)](#), who used data for Local Planning Areas in England on the refusal rate

⁴¹ In a sample of homes in Florida, [Ihlanfeldt \(2009\)](#) also found that land use planning increases the desirability of a community, in that a balanced-budget expansion in expenditures on comprehensive land use planning raises house prices.

of major residential projects and the fraction of major residential projects that are delayed over 13 weeks. Another is Solé-Ollé and Viladecans-Marsal (2012) who used data on the amount of new land zoned for development by local municipal governments in Spain.⁴² Also, Fu and Somerville (2001) analyzed height restrictions in Shanghai. The chapter in this volume by Brueckner and Lall discusses the role of regulation in restricting the supply of affordable housing in developing countries. The lack of international evidence on regulation is particularly unfortunate because the United States appears to be on the one end of the spectrum in regard to the elasticity of housing supply. Caldera and Johansson (2013) estimated this elasticity for each of 21 OECD countries using a two-equation error-correction model and found that the United States has the most elastic supply. It also has the lowest average time required to obtain a building permit among these 21 countries. In general, they found that more elastic countries tend to have shorter periods of time required to obtain a permit.⁴³

Due to the difficulty of measuring regulation, theory on this topic tends to be much more advanced than empirical work. Yet there are still some gaps on the theoretical side as well. One such area is on the effects of developers and other prodevelopment participants in the local political process. Whereas a number of researchers have proposed models to explain the role of homeowners in the formation of regulation, much less attention has been devoted to the opposing forces. Another area that seems worthy of further theoretical work is the effect of regulation on homeownership. Many policy-makers denounce housing supply regulation as an impediment to homeownership because it raises house prices. Yet simple theories of regulation would predict proportional increases in house prices and rents, leaving the incentives for homeownership unchanged. In fact, regulation might promote homeownership if it reduces the likelihood of nearby development that would negatively affect local house prices. We are not aware of any formal models on this topic, and such theory would be helpful in informing policy discussions.

In this chapter, we have also highlighted a number of empirical puzzles that warrant further exploration. One is the lack of correlation between regulation and homeownership given the many models that predict homeowners to be strong supporters of regulation. Another dissonance in the literature is the correlation of regulation with past land

⁴² Cheshire and Hilber (2008) used the indirect method proposed by Glaeser and Gyourko (2003) to examine supply restrictions for office space in 14 British and 8 continental European locations.

⁴³ A number of other studies have estimated the elasticity of housing supply for a smaller set of countries and compared them to qualitative assessments of the regulatory environment for residential construction. These studies include Vermeulen and Rouwendal (2007) for the Netherlands, Malpezzi and MacLennan (2001) for the United Kingdom and the United States, Malpezzi and Sa-Aadu (1996) for Africa, Bertaud and Malpezzi (2001) for Malaysia, Green et al. (1994) for Korea, and Whitehead et al. (2009) for the United Kingdom, Bulgaria, and China.

use. When comparing locations within a single metropolitan area, it seems that municipalities that were historically more densely built are less regulated today, at least in regard to allowed future density. But in a cross-section of metropolitan areas, locations with greater historical density are *more* regulated. A third is the fact that the most recent house price cycle was not larger in amplitude in markets with more regulation.

The vast majority of papers on the effects of regulation have focused on implications for the housing market, and in this chapter, we have discussed effects on house prices, construction, the elasticity of housing supply, the volatility of house prices and construction, and urban form. We have also discussed effects beyond the housing market that include sorting of households by income and demographic characteristics and effects on local labor markets. Yet regulation could potentially affect many other aspects of the economy as well. A few ideas that seem worth exploring include the effect of regulation on labor productivity and output given that restrictions on the housing supply affect the number and skill distribution of available workers, the effect of regulation on renovation expenditures as it might encourage households to remain in their current home rather than trading up to a newly built home, and the effects of regulation on access to credit by raising house prices and thereby boosting the wealth of homeowners.

We close with a few words about the welfare consequences of regulation. It is not obvious that regulation should reduce welfare, even if it raises housing costs, because many forms of regulation are designed to generate local amenities or mitigate negative externalities. However, most theoretical and empirical research has found that the costs of regulation outweigh the benefits by a substantial margin. Given this result, it might be somewhat puzzling that land use regulations are so ubiquitous. One possibility is that agents who do benefit from regulation have more influence in the local political process, leading to a result that benefits a few at the expense of the majority. Another possibility is that local policymakers and their constituents have an inflated view of the benefits or underestimate the costs. A third possibility is that researchers have been unable to measure the benefits of regulation adequately. Clearly, more research would be helpful to investigate the welfare effects of regulation, especially the distributional consequences and the interaction of these effects with the local political process.⁴⁴ With regulatory constraints on the housing supply continuing to proliferate and become stricter over time, further research on all of the causes and consequences of these regulations is only becoming more important for improving our understanding of housing markets, local public finance, and the distribution of households and economic activity across space.

⁴⁴ Cheshire and Sheppard (2002) examined distributional effects and found that regulation has a positive net cost for most of the households in their sample. But their sample is small and consists only of homeowners.

ACKNOWLEDGMENTS

Will Strange and the other editors provided helpful advice on this chapter. We also are grateful to Holger Seig for his comments on an earlier version. Naturally, we remain responsible for any errors or omissions. J. G. thanks the Research Sponsors Program of the Samuel Zell and Robert Lurie Real Estate Center at Wharton for financial support. The analysis and conclusions set forth are those of the authors and do not indicate concurrence by other members of the research staff or the Board of Governors.

REFERENCES

- Baldassare, M., Protash, W., 1982. Growth controls, population growth, and community satisfaction. *Am. Sociol. Rev.* 47 (3), 339–346.
- Barseghyan, L., Coate, S., 2013. Property Taxation, Zoning, and Efficiency: A Dynamic Analysis. Cornell University, Revised September 2013.
- Bartik, T., 1991. Who Benefits from State and Local Economic Development Policies? W.E. Upjohn Institute for Employment Research, Kalamazoo, MI.
- Bates, L.J., Santerre, R.E., 1994. The determinants of restrictive residential zoning: some empirical findings. *J. Reg. Sci.* 34 (2), 253–263.
- Bertaud, A., Brueckner, J.K., 2005. Analyzing building-height restrictions: predicted impacts and welfare costs. *Reg. Sci. Urban Econ.* 25, 109–125.
- Bertaud, A., Malpezzi, S., 2001. Measuring the costs and benefits of urban land use regulation: a simple model with an application to Malaysia. *J. Hous. Econ.* 10, 393–418.
- Black, J.T., Hoben, J.E., 1985. Land price inflation and affordable housing. *Urban Geogr.* 6, 27–47.
- Blanchard, O., Katz, L.F., 1992. Regional evolutions. *Brook. Pap. Econ. Act.* 1992, 1–61.
- Brooks, L., Lutz, B., 2012. From today's city to tomorrow's city: an empirical investigation of urban land assembly. Working paper.
- Brooks, L., Lutz, B., 2014. Vestiges of transit: urban persistence at a micro scale. Working paper.
- Brueckner, J.K., 1990. Growth controls and land values in an open city. *Land Econ.* 66 (3), 237–248.
- Brueckner, J.K., 1995. Strategic control of growth in a system of cities. *J. Public Econ.* 57, 393–416.
- Brueckner, J.K., 1997. Infrastructure financing and urban development: the economics of impact fees. *AEA Pap. Proc.* 66, 383–407.
- Brueckner, J.K., 1998. Testing for strategic interaction among local governments: the case of growth controls. *J. Urban Econ.* 44, 438–467.
- Brueckner, J.K., 2009. Government land use interventions: an economic analysis. In: Lall, S.V., Freire, M., Yuen, B., Rajack, R., Helluin, J.-J. (Eds.), *Urban Land Markets: Improving Land Management for Successful Urbanization*. Springer, New York, NY, pp. 3–23.
- Burge, G., Ihlanfeldt, K., 2006a. Impact fees and single-family home construction. *J. Urban Econ.* 60, 284–306.
- Burge, G., Ihlanfeldt, K., 2006b. The effects of impact fees on multifamily housing construction. *J. Reg. Sci.* 46 (1), 5–23.
- Calabrese, S., Epple, D., Romano, R., 2007. On the political economy of zoning. *J. Public Econ.* 91, 25–49.
- Caldera, A., Johansson, A., 2013. The price responsiveness of housing supply in OECD countries. *J. Hous. Econ.* 22, 231–249.
- Cervero, R., Duncan, M., 2004. Neighbourhood composition and residential land prices: does exclusion raise or lower value? *Urban Stud.* 41 (2), 299–315.
- Chakraborty, A., Knapp, G.-J., Nguyen, D., Shin, J.H., 2010. The effects of high-density zoning on multifamily housing construction in the suburbs of six U.S. metropolitan areas. *Urban Stud.* 47 (2), 437–451.
- Cheshire, P.C., Hilber, C.A.L., 2008. Office space supply restrictions in Britain: the political economy of market revenge. *Econ. J.* 118, 185–221.
- Cheshire, P., Sheppard, S., 2002. The welfare economics of land use planning. *J. Urban Econ.* 52, 242–269.

- Clark, T.N., Goetz, E.G., 1994. The antigrowth machine. In: Clark, T.N. (Ed.), *Urban Innovation: Creative Strategies for Turbulent Times*. SAGE Publications, Thousand Oaks, CA, pp. 105–145.
- Colwell, P., Kau, J., 1982. The economics of building codes and standards. In: Bruce Johnson, M. (Ed.), *Resolving the Housing Crisis*. Pacific Institute for Public Policy Research, San Francisco, CA.
- Cunningham, C., 2007. Growth controls, real options and land development. *Rev. Econ. Stat.* 89 (2), 343–358.
- Davidoff, T., 2013. Supply elasticity and the housing cycle of the 2000s. *Real Estate Econ.* 41 (4), 793–813.
- Davis, M., Heathcote, J., 2004. The price and quantity of residential land in the United States. Discussion Paper No. 2004–37. Board of Governors of the Federal Reserve System, Finance and Economics, Washington, DC.
- Davis, M.A., Palumbo, M.G., 2008. The price of residential land in large US cities. *J. Urban Econ.* 63, 352–384.
- Donovan, T., Neiman, M., 1995. Local growth control policy and changes in community characteristics. *Soc. Sci. Q.* 76 (4), 780–793.
- Downs, A., 1973. *Opening Up the Suburbs: An Urban Strategy for America*. Yale University Press, New Haven, CT.
- Dubin, J.A., Roderick Kiewiet, D., Noussair, C., 1992. Voting on growth control measures: preferences and strategies. *Econ. Polit.* 4, 191–213.
- Ellickson, R., 1977. Suburban growth controls: an economic and legal analysis. *Yale Law J.* 86, 385–511.
- Engle, R., Navarro, P., Carson, R., 1992. On the theory of growth controls. *J. Urban Econ.* 32, 269–283.
- Epple, D., Romer, T., Filimon, R., 1988. Community development with endogenous land use controls. *J. Public Econ.* 35, 133–162.
- Evenson, B., Wheaton, W.C., 2002. Why Local Governments Impose Land-Use Regulations, *Mimeo*.
- Evenson, B., Wheaton, W.C., 2003. Local variation in land use regulations [with comments]. *Brookings-Wharton Papers on Urban Affairs*: 2013, pp. 221–260.
- Fischel, W.A., 1985. *The Economics of Zoning Laws: A Property Rights Approach to American Land Use Controls*. John Hopkins University Press, Baltimore, MD.
- Fischel, W.A., 2001. *The Homevoter Hypothesis*. Harvard University Press, Cambridge, MA.
- Fischel, W.A., 2004. An economic history of zoning and a cure for its exclusionary effects. *Urban Stud.* 41 (2), 317–340.
- Fischel, W.A., 2008. Political structure and exclusionary zoning: are small suburbs the big problem? In: Ingram, G.K., Hong, Y.-H. (Eds.), *Fiscal Decentralization and Land Policies*. Lincoln Institute of Land Policy, Cambridge, MA, pp. 111–136.
- Foster, D., Summers, A.A., 2005. Current state legislative and judicial profiles on land-use regulations in the U.S. Wharton Real Estate Center Working paper No. 512. The Wharton School, University of Pennsylvania, Philadelphia, PA.
- Frieden, B., 1979. A new regulation comes to suburbia. *Public Int.* 55, 15–27.
- Fu, Y., Somerville, C.T., 2001. Site density restrictions: measurement and empirical analysis. *J. Urban Econ.* 49, 404–423.
- Ganong, P., Shoag, D., 2013. Why has regional income convergence in the U.S. declined? HKS Working paper No. RWP12-028. Available from SSRN: <http://ssrn.com/abstract=2081216>.
- Geshkov, M.V., DeSalvo, J.S., 2012. The effect of land-use controls on the spatial size of U.S. urbanized areas. *J. Reg. Sci.* 52 (4), 648–675.
- Glaeser, E.L., Gyourko, J., 2003. The impact of building restrictions on housing affordability. *FRBNY Econ. Pol. Rev.* 9, 21–39.
- Glaeser, E., Ward, B., 2009. The causes and consequences of land use regulation: evidence from greater Boston. *J. Urban Econ.* 65 (3), 265–278.
- Glaeser, E.L., Gyourko, J., Saks, R.E., 2005. Why is Manhattan so expensive? Regulation and the rise in housing prices. *J. Law Econ.* 48 (2), 331–369.
- Glaeser, E.L., Gyourko, J., Saks, R.E., 2006a. Urban growth and housing supply. *J. Econ. Geogr.* 6, 71–89.

- Glaeser, E., Schuetz, J., Ward, B., 2006b. Regulation and the Rise of Housing Prices in Greater Boston. Pioneer Institute for Public Policy Research, Boston, MA, Policy Brief PB-2006-1.
- Glaeser, E.L., Gyourko, J., Saiz, A., 2008. Housing supply and housing bubbles. *J. Urban Econ.* 64, 198–217.
- Glickfeld, M., Levine, N., 1992. Regional Growth—Local Reaction: The Enactment and Effects of Local Growth Control and Management Measures in California. Lincoln Institute of Land Policy, Cambridge, MA.
- Green, R.K., Malpezzi, S., Vandell, K., 1994. Urban regulations and the price of land and housing in Korea. *J. Hous. Econ.* 3, 330–356.
- Green, R.K., Malpezzi, S., Mayo, S.K., 2005. Metropolitan-specific estimates of the price elasticity of supply of housing, and their sources. *AEA Pap. Proc.* 95 (2), 334–339.
- Gyourko, J., Saiz, A., 2006. Construction costs and the supply of housing structure. *J. Reg. Sci.* 46 (6), 627–660.
- Gyourko, J., Voith, R., 2002. Capitalization of federal taxes, the relative price of housing, and urban form: density and sorting effects. *Reg. Sci. Urban Econ.* 32, 673–690.
- Gyourko, J., Saiz, A., Summers, A.A., 2008. A new measure of the local regulatory environment for housing markets: the Wharton Residential Land Use Regulatory Index. *Urban Stud.* 45 (3), 693–721.
- Gyourko, J., Mayer, C., Sinai, T., 2013. Superstar cities. *Am. Econ. J. Econ. Pol.* 5 (4), 167–199.
- Hamilton, B., 1978. Zoning and the exercise of monopoly power. *J. Urban Econ.* 5, 116–130.
- Harter-Dreiman, M., 2004. Drawing inferences about housing supply elasticity from house price responses to income shocks. *J. Urban Econ.* 55, 316–337.
- Helsley, R.W., Strange, W.C., 1995. Strategic growth controls. *Reg. Sci. Urban Econ.* 25, 435–460.
- Hilber, C.A.L., Robert-Nicoud, F., 2013. On the origins of land use regulations: theory and evidence from us metro areas. *J. Urban Econ.* 75, 29–43.
- Hilber, C.A.L., Vermeulen, W., 2013. The Impact of Supply Constraints on House Prices in England. Version: July 24, 2013.
- Hilber, C.A.L., Rouwendal, J., Vermeulen, W., 2014. Local economic conditions and the nature of new housing supply. Working paper.
- Ihlanfeldt, K.R., 2004. Introduction: exclusionary land-use regulations. *Urban Stud.* 41 (2), 255–259.
- Ihlanfeldt, K.R., 2007. The effect of land use regulation on housing and land prices. *J. Urban Econ.* 61, 420–435.
- Ihlanfeldt, K.R., 2009. Does comprehensive land-use planning improve cities? *Land Econ.* 85 (1), 74–86.
- Jackson, K., 2014. Do land use regulations stifle residential development? Evidence from California cities. *J. Urban Econ.* Working paper.
- Kahn, M.E., Vaughn, R., Zasloff, J., 2010. The housing market effects of discrete land use regulations: evidence from the California coastal boundary zone. *J. Hous. Econ.* 19, 269–279.
- Katz, L., Rosen, K., 1987. The interjurisdictional effects of growth controls on housing prices. *J. Law Econ.* 30, 149–160.
- Lenon, M.J., Chattopadhyay, S.K., Heffley, D.R., 1996. Zoning and fiscal interdependencies. *J. Real Estate Finance Econ.* 12, 221–234.
- Levine, N., 1999. The effects of local growth controls on regional housing production and population redistribution in California. *Urban Stud.* 36 (12), 2047–2068.
- Libecap, G.D., Lueck, D., 2011. The demarcation of land and the role of coordination property institutions. *J. Polit. Econ.* 119 (3), 426–467.
- Linneman, P., Summers, A.A., Brooks, N., Buist, H., 1990. The state of local growth management. Wharton Real Estate Center Working paper No. 81. The Wharton School, University of Pennsylvania, Philadelphia, PA.
- Logan, J.R., Zhou, M., 1990. The adoption of growth controls in suburban communities. *Soc. Sci. Q.* 71, 118–129.
- Lutz, B., 2013. Quasi-experimental evidence on the connection between property taxes and residential capital investment. *Am. Econ. J. Econ. Pol.*, Revised January 13, 2014, forthcoming.
- Lynch, A.K., Rasmussen, D.W., 2004. Proximity, neighbourhood and the efficacy of exclusion. *Urban Stud.* 41 (2), 285–298.

- Malpezzi, S., 1996. Housing prices, externalities, and regulation in U.S. metropolitan areas. *J. Hous. Res.* 7, 209–241.
- Malpezzi, S., MacLennan, D., 2001. The long-run price elasticity of supply of new residential construction in the United States and the United Kingdom. *J. Hous. Econ.* 10, 278–306.
- Malpezzi, S., Sa-Aadu, J., 1996. What have African housing policies wrought? *Real Estate Econ.* 24 (2), 133–160.
- Malpezzi, S., Wachter, S.M., 2005. The role of speculation in real estate cycles. *J. Real Estate Lit.* 13 (2), 143–164.
- Mayer, C.J., Somerville, C.T., 2000. Land use regulation and new construction. *Reg. Sci. Urban Econ.* 30, 639–662.
- McDonald, J.F., 1995. Houston remains unzoned. *Land Econ.* 71 (1), 137–140.
- McDonald, J.F., McMillen, D.P., 2004. Determinants of suburban development controls: a Fischel expedition. *Urban Stud.* 41, 341–361.
- McFarlane, A., 1999. Taxes, fees, and urban development. *J. Urban Econ.* 46, 416–436.
- McMillen, D.P., McDonald, J.F., 1991. Urban land value function with endogenous zoning. *J. Urban Econ.* 29, 14–27.
- Mills, E.S., 2005. Why do we have urban density controls? *Real Estate Econ.* 33 (3), 571–585.
- Molotch, H., 1976. The city as a growth machine: toward a political economy of place. *Am. J. Sociol.* 82 (2), 309–332.
- Neiman, M., 1980. Zoning policy, income clustering and suburban change. *Soc. Sci. Q.* 61, 666–675.
- Nelson, A.C., Dawkins, C.J., Sanchez, T.W., 2004. Urban containment and residential segregation: a preliminary investigation. *Urban Stud.* 41 (2), 423–439.
- Nichols, J.B., Oliner, S.D., Mulhall, M.R., 2013. Swings in commercial and residential land prices in the United States. *J. Urban Econ.* 72, 57–76.
- Noam, E.M., 1983. The interaction of building codes and housing prices. *AREUEA J.* 10, 394–404.
- Ortalo-Magne, F., Prat, A., 2014. On the political economy of urban growth: homeownership versus affordability. *Am. Econ. J. Microecon.* 6 (1), 154–181.
- Oster, S., Quigley, J., 1977. Regulatory barriers of the diffusion of innovation: some evidence from building codes. *Bell J. Econ.* 8, 361–377.
- Paciorek, A., 2013. Supply constraints and housing market dynamics. *J. Urban Econ.* 77, 11–29.
- Pasha, H.A., 1992. Comparative statics analysis of density controls. *J. Urban Econ.* 32, 284–298.
- Pendall, R., 2000. Local land use regulation and the chain of exclusion. *APA J.* 66 (2), 125–142.
- Pendall, R., Puentes, R., Martin, J., 2006. From Traditional to Reformed: A Review of Land Use Regulations in the Nation's 50 Largest Metropolitan Areas. Metropolitan Policy Program, The Brookings Institution, Washington, DC.
- Pogodzinski, J.M., Sass, T.R., 1994. The theory and estimation of endogenous zoning. *Reg. Sci. Urban Econ.* 24, 601–630.
- Pollakowski, H.O., Wachter, S.M., 1990. The effects of land-use constraints on housing prices. *Land Econ.* 66 (3), 315–324.
- Quigley, J.M., Raphael, S., 2005. Regulation and the high cost of housing in California. *AEA Pap. Proc.* 95 (2), 323–328.
- Quigley, J., Rosenthal, L., 2005. The effects of land use regulation on the price of housing: what do we know? What can we learn? *Cityscape* 81 (1), 69–137.
- Quigley, J.M., Swoboda, A.M., 2007. The urban impacts of the endangered species act: a general equilibrium approach. *J. Urban Econ.* 61, 299–318.
- Quigley, J.M., Raphael, S., Rosenthal, L.A., 2004. Local land-use controls and demographic outcomes in a booming economy. *Urban Stud.* 41 (2), 389–421.
- Rolleston, B.S., 1987. Determinants of restrictive suburban zoning: an empirical analysis. *J. Urban Econ.* 21, 1–21.
- Saiz, A., 2010. The geographic determinants of housing supply. *Q. J. Econ.* 125 (3), 1253–1296.
- Saks, R.E., 2008. Job creation and housing construction: constraints on metropolitan area employment growth. *J. Urban Econ.* 62, 187–195.

- Segal, D., Srinivasan, P., 1985. The impact of suburban growth restrictions on U.S. housing price inflation, 1975–1978. *Urban Geogr.* 6 (1), 14–26.
- Shertzer, A., Twinam, T., Walsh, R.P., 2014. Race, ethnicity, and discriminatory zoning. NBER Working paper Series. Working paper 20108 (May). <http://www.nber.org/papers/w20108>.
- Shlay, A.B., Rossi, P.H., 1981. Keeping up the neighborhood: estimating net effects of zoning. *Am. Sociol. Rev.* 46, 703–719.
- Skidmore, M., Peddle, M., 1998. Do development impact fees reduce the rate of residential development? *Growth Change* 29, 383–400.
- Solé-Ollé, A., Viladecans-Marsal, E., 2012. Lobbying, political competition, and local land supply: recent evidence from Spain. *J. Public Econ.* 96, 10–19.
- Song, Y., Zenou, Y., 2006. Property tax and urban sprawl: theory and implications for US cities. *J. Urban Econ.* 60, 519–534.
- Sonstelie, J.C., Portney, P.R., 1978. Profit maximizing communities and the theory of local public expenditures. *J. Urban Econ.* 5, 263–277.
- Thorson, J.A., 1997. The effect of zoning on housing construction. *J. Hous. Econ.* 6, 81–91.
- Turner, M.A., Haughwout, A., van der Klaauw, W., 2014. Land use regulation and welfare. *Econometrica* 82, 1341–1403, Revised November 2013.
- Vermeulen, W., Rouwendal, J., 2007. Housing supply and land use regulation in the Netherlands. Tinbergen Institute Discussion Paper. Version: July 2007. Available from: <http://ssrn.com/abstract=1003955>.
- Wallace, N.E., 1988. The market effects of zoning undeveloped land: does zoning follow the market? *J. Urban Econ.* 23, 307–326.
- Wheaton, W.C., Chervachidze, S., Nechayev, G., 2014. Error Correction Models of MSA Housing ‘Supply’ Elasticities: Implications for Price Recovery, 3rd Draft: March 9, 2014.
- Whitehead, C., Chiu, R.L.H., Tsenkova, S., Turner, B., 2009. Land use regulation: transferring lessons from developed economies. In: Lall, S.V., Freire, M., Yuen, B., Rajack, R., Helluin, J.-J. (Eds.), *Urban Land Markets: Improving Land Management for Successful Urbanization*. Springer, Dordrecht, pp. 51–69.
- Zabel, J., 2012. Migration, housing market, and labor market responses to employment shocks. *J. Urban Econ.* 72, 267–284.
- Zabel, J., Dalton, M., 2011. The impact of minimum lot size regulations on house prices in Eastern Massachusetts. *Reg. Sci. Urban Econ.* 41, 571–583.
- Zhou, J., McMillen, D.P., McDonald, J.F., 2008. Land values and the 1957 comprehensive amendment to the Chicago zoning ordinance. *Urban Stud.* 45 (8), 1647–1661.

This page intentionally left blank

CHAPTER 20

Transportation Costs and the Spatial Organization of Economic Activity

Stephen J. Redding*, Matthew A. Turner†

*Economics Department and WWS, Princeton University Fisher Hall, Princeton, NJ, USA

†Economics Department, Brown University, Providence, RI, USA

Contents

20.1. Introduction	1340
20.2. Stylized Facts About Transportation	1343
20.2.1 Transportation costs for goods	1343
20.2.2 Household travel and commuting	1352
20.2.3 External costs	1355
20.3. Theoretical Framework	1355
20.3.1 Preferences and endowments	1356
20.3.2 Production technology	1357
20.3.3 Market access and wages	1358
20.3.4 Labor market equilibrium	1359
20.3.5 Trade flows	1360
20.3.6 Welfare	1361
20.3.7 General equilibrium	1364
20.3.8 Counterfactuals	1365
20.4. Reduced-Form Econometric Framework	1366
20.4.1 A simple taxonomy	1366
20.4.2 Identification of causal effects	1367
20.4.2.1 <i>Planned route instrumental variable approach</i>	1368
20.4.2.2 <i>Historical route instrumental variable approach</i>	1369
20.4.2.3 <i>Inconsequential units approach</i>	1369
20.4.3 Distinguishing growth from reorganization	1370
20.5. Reduced-Form Empirical Results	1372
20.5.1 Intracity infrastructure and the geographic organization of economic activity	1372
20.5.1.1 <i>Infrastructure and decentralization</i>	1372
20.5.1.2 <i>Infrastructure and miscellaneous city-level outcomes</i>	1374
20.5.1.3 <i>Subways and the internal organization of cities, and related other results</i>	1376
20.5.2 Intercity infrastructure and the geographic organization of economic activity	1378
20.5.2.1 <i>High-income countries</i>	1378
20.5.2.2 <i>Low-income countries</i>	1380
20.6. Discussion	1383
20.6.1 Growth versus reorganization	1383
20.6.2 The effects of transportation infrastructure on economic activity	1384
20.6.2.1 <i>Invariance across economies</i>	1384

20.6.2.2	<i>Variability across activities and modes</i>	1386
20.6.2.3	<i>Political economy of infrastructure allocation</i>	1387
20.6.3	General equilibrium effects	1387
20.6.4	Structural estimation, general equilibrium, and welfare	1388
20.7.	Conclusion	1393
	Acknowledgments	1394
	References	1394

Abstract

This chapter surveys the theoretical and empirical literature on the relationship between the spatial distribution of economic activity and transportation costs. We develop a multiregion model of economic geography that we use to understand the general equilibrium implications of transportation infrastructure improvements within and between locations for wages, population, trade, and industry composition. Guided by the predictions of this model, we review the empirical literature on the effects of transportation infrastructure improvements on economic development, paying particular attention to the use of exogenous sources of variation in the construction of transportation infrastructure. We examine evidence from different spatial scales, between and within cities. We outline a variety of areas for further research, including distinguishing reallocation from growth and dynamics.

Keywords

Highways, Market access, Railroads, Transportation

JEL Classification Codes

F15, R12, R40

20.1. INTRODUCTION

The organization of economic activity in geographic space depends crucially on the transportation of goods and people. Most production involves the movement of inputs such as raw materials, labor, and fuel from different locations. Most consumption requires either the conveyance of finished goods or the transfer of people to the points at which goods and services are supplied. The transportation sector as a whole typically accounts for around 5% of gross domestic product (GDP), and transportation networks constitute some of the largest investments ever made. In the United States, the Interstate Construction Program extended to 42,795 miles of highways with an estimated cost of \$128.9 billion (1991 US dollars).¹ Multiplying estimates of the cost per interstate lane kilometer found in [Duranton and Turner \(2012\)](#) by the extent of the system gives

¹ US Department of Transportation, Federal Highway Administration, interstate cost estimates reported to Congress.

much larger values. In China, the National Trunk Highway System involved the construction of around 21,747 miles (35,000 km) of highways over a period of 15 years at an estimated construction cost of around \$120 billion (in current price US dollars).²

Transportation technologies themselves have undergone large-scale changes over time, which have in turn reshaped the spatial organization of economic activity. For most of human history, the movement of goods and people was limited by the physical capabilities of humans and their animals. The invention of the railroad reduced transportation costs and created a hub-and-spoke transportation network that was characterized by substantial fixed costs (e.g., in stations and goods yards) and favored point-to-point travel between the central cities. The development of the internal combustion engine (and hence the automobile and truck) in turn created greater flexibility in transportation, benefiting lower-density locations relative to central cities.³ Even within existing transportation technologies, such as maritime shipping, there have been large-scale changes in the organization of economic activity in the form of containerization and the adoption of new information and communication technologies such as the computer. These innovations have played an important part in the development of integrated logistics networks, which control the movement of a package from its origin to its destination, and integrate packaging, storage, transport, inventories, administration, and management. The discovery of entirely new modes of transportation, such as air travel, has further transformed the relative attractiveness of locations for economic activity.

This chapter describes our current understanding of the way that transportation costs and transportation infrastructure affect the organization of economic activity within a country. We first provide some basic facts about transportation costs within and between cities. Then we develop a multiregion model of economic geography as a framework to organize our discussion of the empirical literature. The existing empirical literature on the effects of transportation costs and infrastructure can be usefully divided into two parts. The first of these parts considers the role of transportation costs between cities and is mainly interested in the movement of goods, while the second considers the role of transportation costs within cities and is mainly interested in the movement of people. Our model unifies the analysis of within-city and between-city transportation, thereby allowing us to simultaneously consider the two previously disparate strands of the empirical literature. Analysis of our model yields structural equations corresponding to the reduced-form estimating equations on which the two parts of the empirical literature are based. The divergence between theoretically founded

² Faber (2015).

³ See, for example, the discussion in Glaeser and Ponzetto (2013).

structural equations and reduced-form estimating equations, in turn, provides insight into the inference problems that reduced-form estimation must overcome. Finally, with a handful of exceptions, the existing literature provides only an incomplete understanding of general equilibrium effects of transportation infrastructure and little basis for welfare analysis. The model that we develop illustrates a possible direction for research on this issue.

The available empirical literature provides credible, causal estimates of the effect of roads, railroads, and subways on outcomes such as population density, land rents, and output. In addition to providing particular elasticity estimates, this literature is large enough to suggest three preliminary conclusions. First, that the effects of different types of infrastructure are similar across economies at different stages of development and are not especially sensitive to the spatial scale of the unit of observation. Second, that different modes of transportation are not interchangeable. Railroads affect production more than the population and the effects of railroads on the location of production vary systematically with the weight-to-value ratio of output, while the spatial organization of the population is more sensitive to roads and subways than to railroads. Finally, and unsurprisingly, institutions matter. The existing empirical literature suggests that politics plays an important role in the allocation of infrastructure and that these politics vary systematically across countries.

Determining the extent to which the effects of transportation infrastructure reflect growth or reorganization is fundamental to understanding its role in the spatial organization of economic activity. Indeed, this question is at the heart of Fogel's classic study of railroads in the late nineteenth century United States. While the current empirical literature provides credible causal estimates of the effects of transportation infrastructure, it is impossible for the reduced-form regressions conducted in almost all of the empirical studies that we survey to separately identify the effect of transportation infrastructure on the growth and reorganization of economic activity. We suggest two approaches to this problem: one is a simple extension of the existing reduced-form literature, and the second is an implementation of our structural model. The handful of articles which shed light on this question suggest that reorganization is often about as important as growth. This is an important area for further research.

The remainder of this chapter is structured as follows. [Section 20.2](#) reports some descriptive evidence regarding transportation costs across countries and over time. [Section 20.3](#) introduces the theoretical framework that we use to organize our discussion of the empirical evidence. [Section 20.4](#) uses the model to develop a reduced-form framework for examining the impacts of transportation infrastructure on the distribution of economic activity between and within cities. [Section 20.5](#) uses this reduced-form framework to review existing empirical evidence of these impacts. [Section 20.6](#) discusses the interpretation of this existing evidence. [Section 20.7](#) summarizes our conclusions.

20.2. STYLIZED FACTS ABOUT TRANSPORTATION

In this section, we present stylized facts about transportation costs for goods and people, both over a long historical time period and across countries.⁴ The key features of the data are as follows. First, there is a secular decline in transportation costs for goods. Second, there is a change in the relative importance of different transportation modes over time (e.g., rail versus road versus air) and for value versus weight. Third, transportation costs for people continue to be important. Commuting costs remain substantial, both in terms of the opportunity cost of time and in terms of overall household expenditure.

20.2.1 Transportation costs for goods

To provide a rough indication of the real resources involved in the transportation sector over time, Figure 20.1 displays the share of the transportation sector in US GDP from the late nineteenth century to the late twentieth century.⁵ The striking feature of this figure is the long secular decline in the share of the transportation sector, which is even

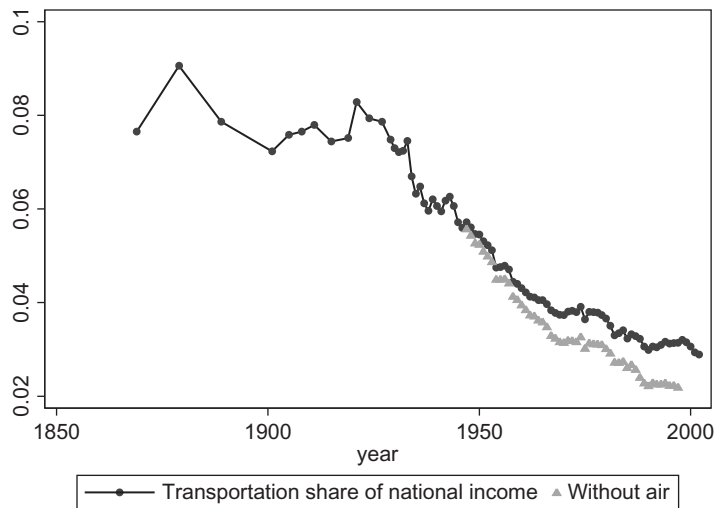


Figure 20.1 Share of transport sector in US GDP. Sources: Department of Commerce (since 1929), and *Historical Statistics of the United States (Martin Series)* before then.

⁴ For a more detailed analysis of the evolution of transportation costs over time in the United States, see Glaeser and Kohlhase (2004).

⁵ Figures 20.1–20.4 replicate similar figures that appear in Glaeser and Kohlhase (2004).

more rapid toward the end of the twentieth century if air transport is excluded. The share of US GDP attributed to transportation⁶ fell from about 8% in 1929 to about 3% by 1990, of which about one-quarter is air transport. While these numbers are striking, they may reflect the increased importance of nontraded services rather than a decrease in the importance of transportation. In addition, while these GDP figures tell us about the resources devoted to moving goods, they do not tell us about the amount and value of the goods being moved.

To provide a more direct measure, Figure 20.2 displays the transportation costs for a given mode of transportation (railroads) in the United States over a similar time period (measured as costs per ton mile in 2001 dollars). The figure confirms a secular decline in transportation costs over time. The price per ton mile of rail freight fell from about 18.5 cents in 1890 to about 2 cents in 2000. Figure 20.3 compares the evolution of truck, rail, and pipeline transportation costs for the United States during the post-Second World War period (measured as revenue per ton mile in 2001 dollars). As is

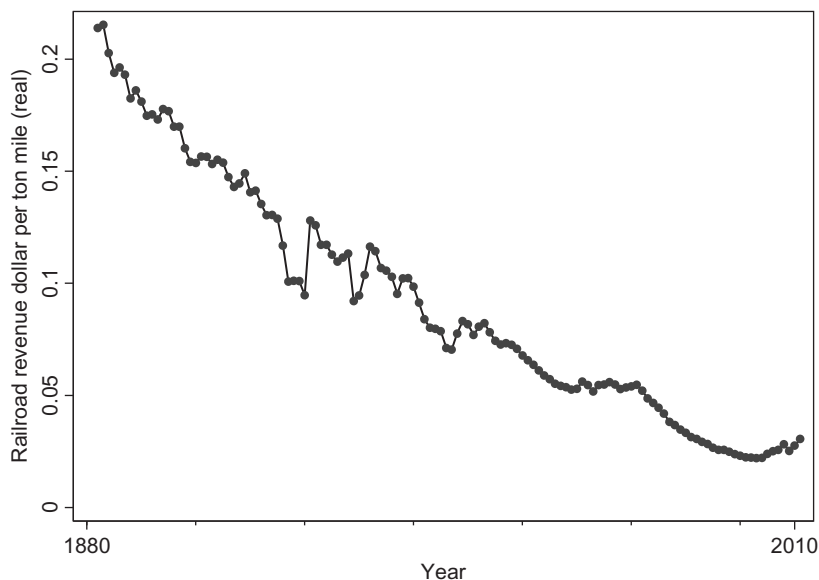


Figure 20.2 The costs of railroad transportation over time. Sources: *Historical Statistics of the United States* (until 1970), and *Bureau of Transportation Statistics annual reports 1994 and 2002*.

⁶ Defined as rail, water, pipeline, trucking, warehousing, air transport, transportation services, and local and interurban rail transit.

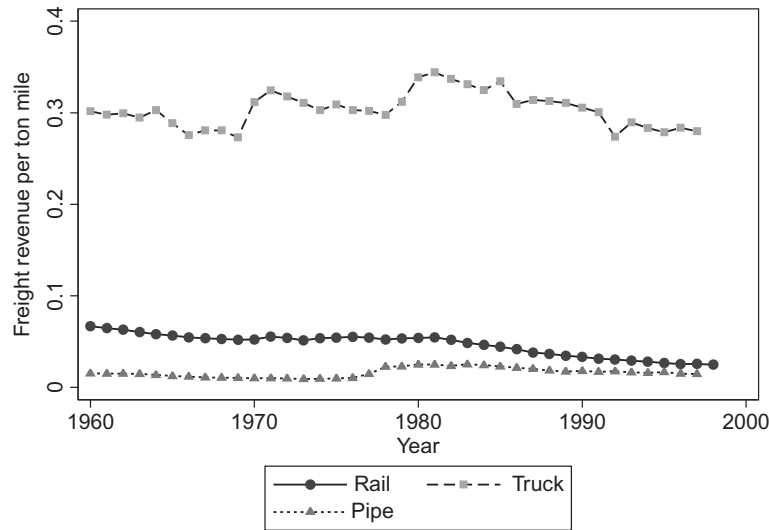


Figure 20.3 Revenue per ton mile (all modes). Sources: Bureau of Transportation Statistics annual reports.

apparent from the figure, truck transport is substantially more expensive than rail transport, and its real costs have fallen notably since the early 1980s.⁷

Figure 20.4 shows the evolution of ton miles of freight over time from the mid-1960s. Rail is relatively more important than trucks when we measure volume shipped rather than value because of a widely observed selection effect in which more expensive items are disproportionately shipped by the more expensive transportation mode.⁸ As a share of the value of goods, Glaeser and Kohlhase (2004) find that for heavy low-value goods traveling by truck (e.g., lumber), the cost of shipping goods over an average shipment distance can be as high as 20% of the value of the good. For more typical sectors, this value is on the order of 5%. For goods traveling by rail, the corresponding values range from 0.1% to 2%. These findings highlight that the cost of moving freight has dropped dramatically to the point that freight transportation is about 3% of the US economy and that freight charges make up only a small share of the value of final output.

⁷ These figures invite the question of why people use trucks at all, the nominally more costly mode. Although trucks have a higher cost per ton mile than rail, the real cost of quality-adjusted transportation services also depends on speed, flexibility, reliability, and a number of other attributes. The large-scale reallocation of transportation expenditure from rail to trucks following the invention of the internal combustion engine suggests that this invention was associated with a substantial reduction in the real cost of quality-adjusted transportation services, at least for many types of shipments and journeys.

⁸ This is an example of the Alchian–Allen effect from the international trade literature or “shipping the good apples out.”

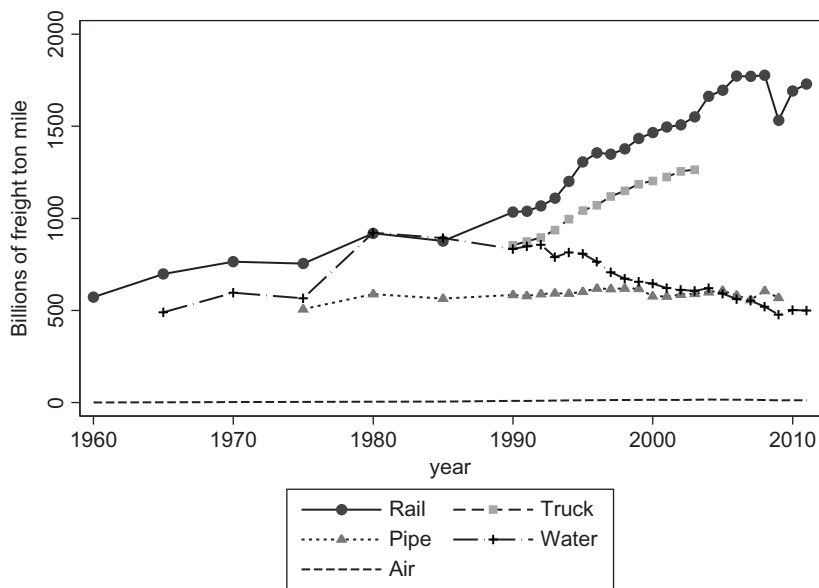
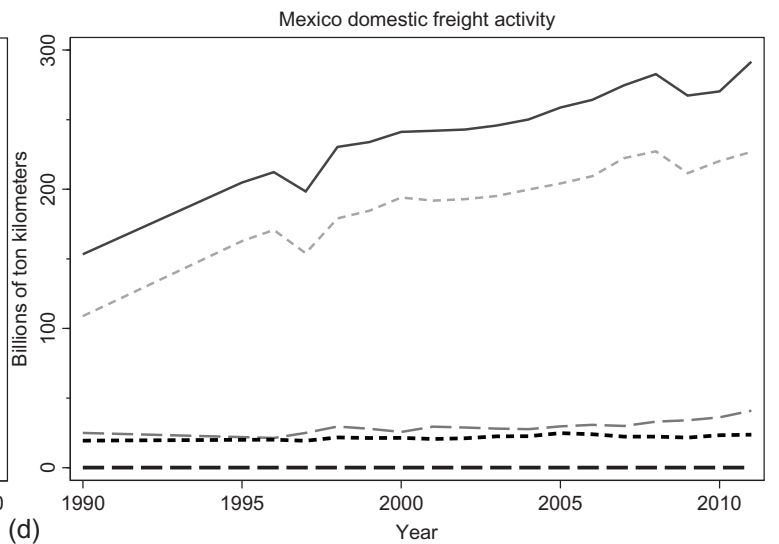
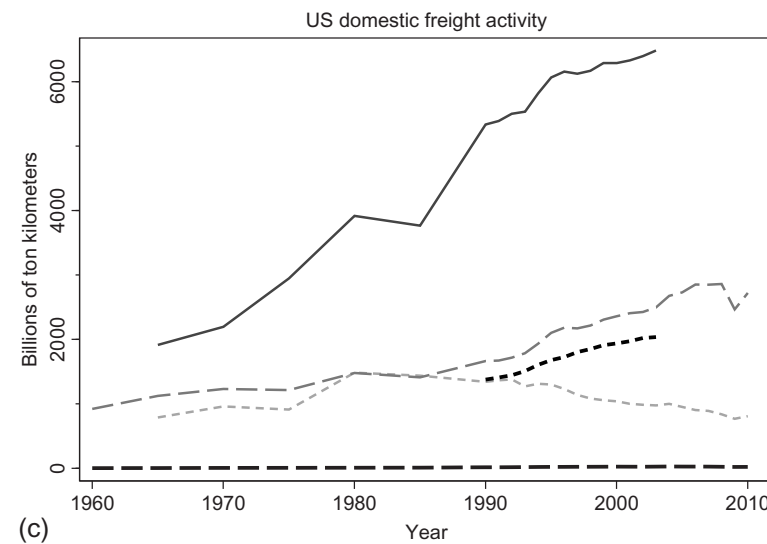
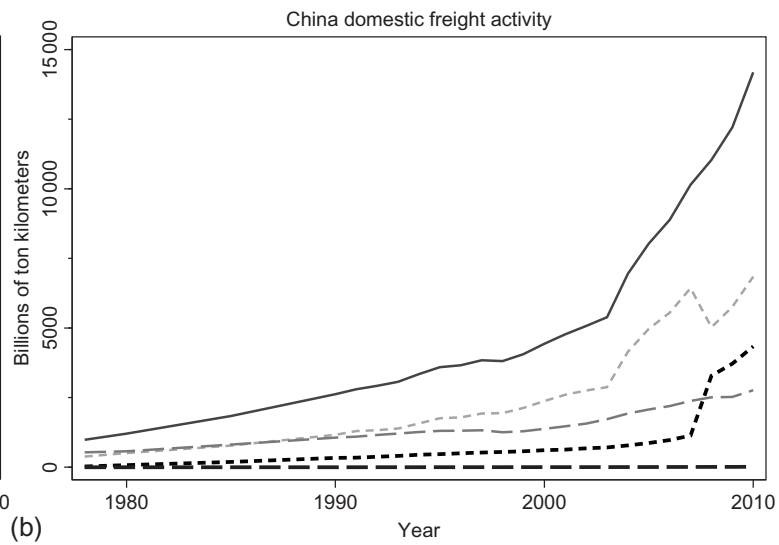
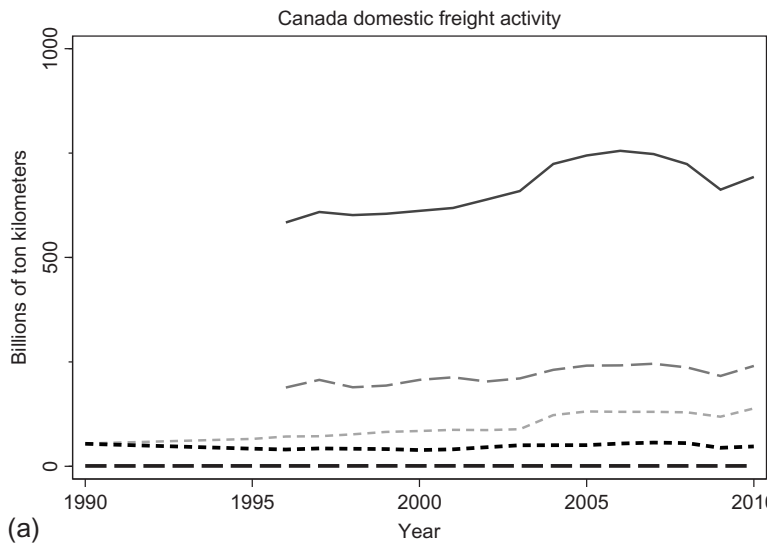


Figure 20.4 Ton miles of freight over time. Sources: *Bureau of Transportation Statistics annual reports.*

To show that these patterns are not specific to the United States, [Figure 20.5](#) describes ton kilometers of domestic trade for seven countries by mode and year. While there are differences between countries, several patterns are clear. First, there is a general trend upward in the amount of domestic trade, as expected given a secular increase in the level economic activity over time. Second, trucking is not the dominant mode of travel in any of our countries by this metric. Third, the amount of material being moved is immense.⁹

[Table 20.1](#) shows the value of international trade by mode for a sample of countries in 2007. In [Figure 20.5](#) we see that the share of ton kilometers for goods that travel by air is negligible. In contrast, in [Table 20.1](#), we see that the share of the value of goods traveling by air is often large. While the two sets of results are not directly comparable (one measures domestic trade and the other measures international trade), together they strongly suggest that high-value goods travel by air and low-value goods travel by ship or rail.

⁹ To get a sense of this, a typical coal train in the United States is about 100 cars long, about 2 km, and each car carries about 100 tons of coal, which implies 10,000 tons per train. If such a train travels 100 km, it provides 1 million ton-km of freight service. To carry 5000 billion ton-km of freight per year, a bit less than the current US annual total, we require about 1200 such trains to operate 24 h per day, 365 days per year at 50 km per h.



(Continued)

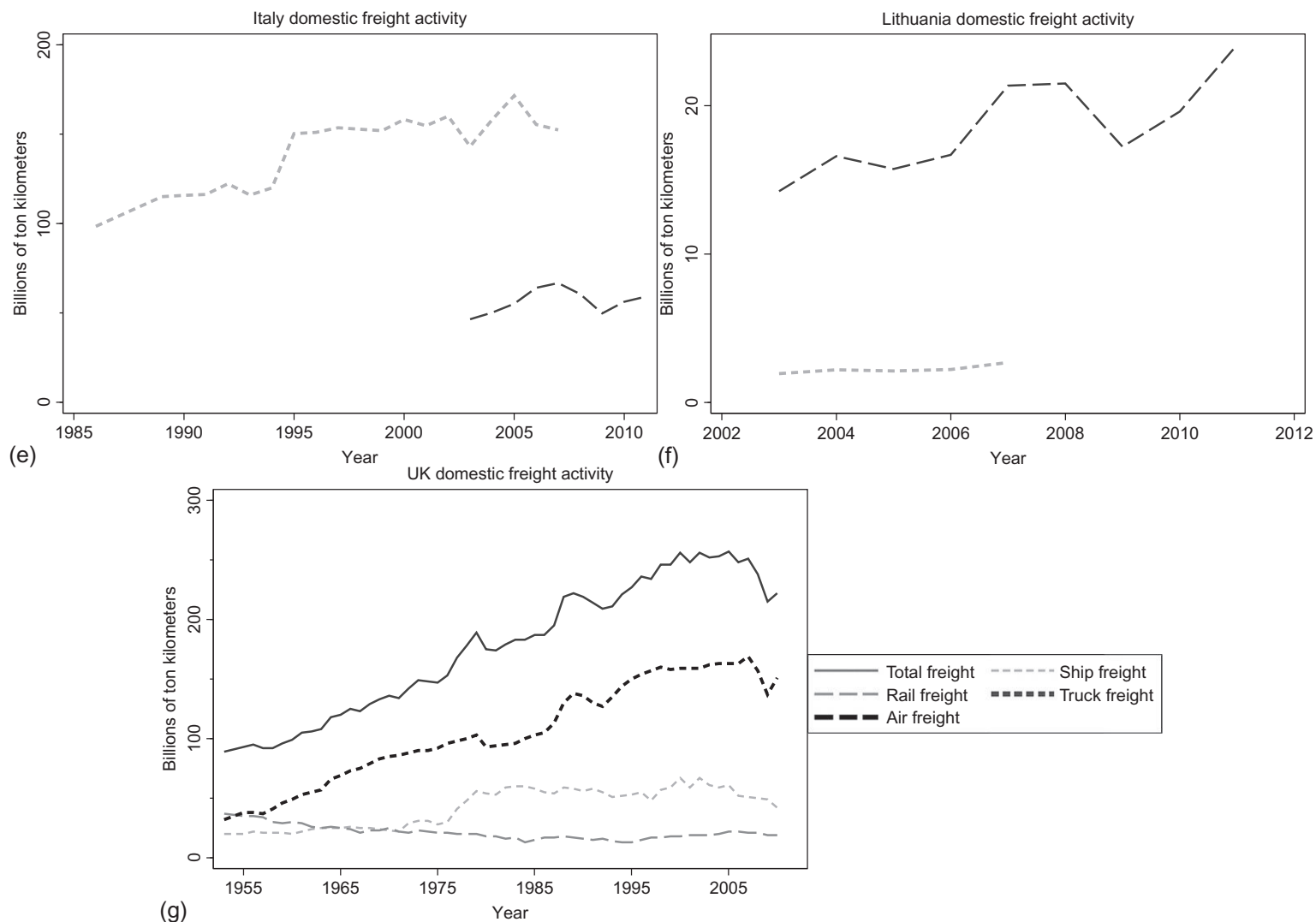


Figure 20.5 Ton kilometers of freight by year and mode for several countries. Sources: (a) *North American Transportation Statistics (2012c)*; (b) *China Data Online (2010)*; (c) *Bureau of Transportation Statistics (2012a)*; (d) *North American Transportation Statistics (2012c)*; (e) *Eurostat (2010)*; (f) *Eurostat (2010)*; (g) *Eurostat (2010)*.

Table 20.1 Shares of total international trade by country and mode

	Total	Ship (%)	Air (%)	Truck (%)	Rail (%)
Austria	279	13	8	70	7
Belgium	750	26	14	44	2
Bulgaria	48	38	4	50	7
Canada	798	19	11	47	12
Cyprus	10	83	16	0	
Czech Republic	194	2	5	82	9
Denmark	186	24	8	11	0
Estonia	24	34	4	34	10
EU27	10,261	29	12	32	2
Finland	169	69	15	12	4
France	1105	25	11	9	0
Germany	2231	27	14	37	3
Greece	106	57	7	33	1
Hungary	164	1	10	75	9
Ireland	198	45	31	11	0
Italy	1003	40	9	36	3
Latvia	21	18	3	45	14
Lithuania	38	29	3	56	12
Mexico	554	24	6	56	8
Netherlands	995	27	9	39	1
Poland	259	17	3	55	9
Portugal	119	37	7	52	0
Romania	101	25	2	13	3
Slovakia	96	13	4	60	12
Slovenia	56	16	3	64	4
Spain	615	45	7	44	2
Sweden	313	25	10	15	1
United Kingdom					
United States	3116	45	25	18	4

Total trade for European countries is in billions of euros and for the other countries is in billions of US dollars.

Source: [North American Transportation Statistics \(2012d\)](#) and [Eurostat \(2012\)](#).

[Table 20.2](#) compares employment in for-hire transportation by mode for Canada, Mexico, and the United States in 2002 or 2003 (depending on data availability). Transportation is typically smaller as a share of employment than as a share of GDP. The share of employment in transportation is about 3% for the United States and Canada, and almost 6% for Mexico. In all three countries, the largest fraction of transportation employment is devoted to trucking. Note that the share of labor devoted to for-hire transportation is close to the same as the share devoted to commuting.

A striking feature of international trade in goods is the extent to which the volume of trade in goods declines with distance. [Hillberry and Hummels \(2008\)](#) examine the

Table 20.2 Employment in for-hire transportation as share of total employment

Country	Year	Total employment	All transport	Air	Rail	Truck	Ship
Canada	2002	13	3.56	0.44	0.28	1.29	0.09
Mexico	2003	35	5.75	0.06	0.04	2.75	0.03
United States	2003	130	3.22	0.41	0.17	1.02	0.04

Total employment is in millions of people, and all others are percentages of the total.

Sources: [North American Transportation Statistics \(2012a\)](#) and [Bureau of Transportation Statistics \(2012b\)](#).

pattern of shipments between US mining, manufacturing, and wholesaling firms and find that three-quarters of all shipments, weighted by the value of shipments, begin and end in the same zip code, a conclusion that does not appear to be driven by shipments from wholesalers to retailers. [Hummels \(1999\)](#) documents the cost of air freight between the 1950s and the first decade of the twenty-first century and finds that it decreases by a factor of about 12.5, while the cost of shipping was approximately constant. For comparison, the corresponding decrease for rail, from [Glaeser and Kohlhase \(2004\)](#), is about a factor of 8 for 110 years.¹⁰ Over the same 1955–2004 period, [Hummels \(1999\)](#) documents 5–7% increases in the value and weight of international trade and an 11% average annual increase in the share of the value of trade that travels by air. [Lima and Venables \(2001\)](#) use data describing market price to ship a standard 40 ft container from Baltimore, Maryland, to 1 of about 50 countries around the world in the late 1990s.¹¹ In a regression of total freight charge on a land-locked country indicator, sea distance and land distance to destination, they find that the cost to ship a standard container 1000 km by sea is about \$190, while to ship it the same distance over land the cost is about \$1380. Recalling that a standard container can hold about 30 tons, we find that this gives sea rates of about half a cent per ton mile and land rates of about 5 cents per ton mile, so overland travel is about 10× as expensive as sea travel. These rates seem somewhat low compared with the price of US truck and rail rates reported in [Glaeser and Kohlhase \(2004\)](#) (28 cents per ton mile for trucks and 3 cents per ton mile for rail). Finally, [Clark et al. \(2004\)](#) find that the cost of shipping all maritime freight to and from the United States is equal to about 5.25% of the value of freight and that port efficiency is an important contributor to this cost.

These facts paint a subtle picture. While the real costs of moving goods have fallen to astonishingly low levels and the weight of trade is immense, the fact that not all

¹⁰ Note that rail was at its peak in 1890 and airfreight was novel in 1950.

¹¹ To get a sense of the nature of the sample, when the countries are ranked by kilometers of paved road per person, the median country is Kenya. For evidence regarding the role of containerization in reducing international transportation costs, see [Bernhofen et al. \(2013\)](#).

trade travels by the cheapest mode and that most trade travels very short distances suggests that the decline in the price per ton of moving goods is not leading to the “death of distance.”

While it is natural to think of time costs as being most important for the movement of people, the rise of air trade suggests that the time in transit is an increasingly important part of the cost of transit for goods. A back of the envelope calculation bolsters this idea. The capacity of a typical 40 ft container is about 30 tons. From [Duranton et al. \(2014\)](#), the value per ton of an average US domestic shipment of electrical appliances is about \$6000 per ton. Thus, a typical container of US electrical appliances can hold about \$200,000 worth of freight. From [Glaeser and Kohlhase \(2004\)](#), shipping this container 1000 miles by rail will cost about $1000 \times 30 \times 0.023 = \$700$. At a 5% annual rate, daily interest on 1 million dollar cargo is $200,000 \times 0.05/365 = \$28$, so on a 5-day journey, the opportunity cost of travel time is equal to about one-fifth of the freight charges. An average ton of goods is worth less than one-tenth of this, while a typical ton of computer equipment is $15\times$ as valuable. At least for relatively high value to weight products, time in transit is important.

Moreover, the predominance of short-haul trade suggests that not only are transportation costs important, but that the geography of production is influenced by transportation costs. For example, the development of nineteenth century Chicago was heavily influenced by its location relative to its surrounding agricultural hinterland, as discussed in [Cronon \(1991\)](#). This points to an important econometric problem in interpreting the transportation cost data presented so far: these data describe equilibrium transportation costs. Therefore, they do not isolate the supply-side production function (or cost function) for transportation, but are rather influenced by both demand and supply. Although these data on transportation costs are still suggestive, they capture both the cost of transportation (supply) and the endogenous organization of economic activity in space in response to the cost of transportation (demand). This presents important and difficult econometric problems, to which we return below.

Another striking feature of microdata on trade and production is the finding of [Atalay et al. \(2013\)](#) that most vertically integrated firms actually ship very little between plants. From the above, we have the puzzling collection of facts: the cost of moving goods is a small fraction of their value, most shipments occur over very small distances, most shipments do not travel by the cheapest mode, and the time cost of freight is probably important. One possible way of rationalizing this combination of findings is that there is something valuable about proximity other than the reduction in transportation costs—that is, agglomeration effects including knowledge spillovers and idea flows. In this case, trade could decline rapidly with distance even in a world in which transportation costs are small, because most economic activity is clustered together for these other reasons and hence most economic interactions are over short distances.

Alternatively, one could question whether transportation costs are really as small as a share of value added as some of the figures above suggest. Arguably, labor used in transportation should be compared with labor used in production and we should take into account the same kinds of costs that we think about for commuting: time costs and scheduling costs.

20.2.2 Household travel and commuting

While the trade literature has typically focused on the movement of goods, another important source of transportation costs in the urban literature is the movement of people. These costs of transporting people remain substantial, both in terms of the opportunity cost of time and in terms of the share of overall household expenditure. Table 20.3 lists average round-trip commuting times in minutes in a sample of countries and years for which data were readily available. While we should be concerned that differences in commuting times across countries reflect sampling error and differences in survey methods, with this caveat, these data suggest that the country mean round-trip commute is about 40 min in the 2000–2005 window, where we have the most observations. These times are fairly closely clustered, with a standard deviation of just less than 8 min. If the

Table 20.3 Average round-trip commuting times in minutes, by country and year

	1980–1984	1990–1994	2005–2009	2010–2012	
Austria			35		
Belgium			41		
Canada		54	59	63	52
Denmark			38		
Finland			41		
France			36		
Germany			42		
Greece			32		
Ireland			36		
Italy			30		
Luxembourg			35		
Netherlands			47		
Portugal			29		
Spain			38		
Sweden			38		
United Kingdom			37	53	
United States	43	45	51		50

Sources: Statistics Canada (2010), Turcotte (2005), Eurofound (2000), Office of National Statistics (2009), and Census Bureau (2009).

“work day” consists of 8 h at work and time in commuting, then commuting consumes about 7.5% of labor. Alternatively, if we value time in commuting at half the wage (as is common in the transportation economics literature; see [Small and Verhoef, 2007](#)) and suppose an 8-h work day, then the value of commuting time is equal to about 3.5% of the value of labor. While this is a large number, it understates the cost of household travel by restricting attention to commuters and commuting trips.

Alternatively, [Schafer \(2000\)](#) summarizes 26 national household travel surveys from countries all over the world. Averaging across these surveys, again with the caveat about the comparability of surveys, he finds that the daily household travel time is about 73 min, with a standard deviation of about 12 min. If we value this time at half the wage and again suppose an 8-h work day, then the value of time spent in household travel is about 8% of the value of labor.¹² If we take the labor share of GDP to be close to the current US level at 0.6, then the time cost of household travel is between 2.4% and 4.8% of GDP.

[Table 20.4](#) describes the shares of household expenditure on transportation for 26 countries and several years. Again noting the possibility of different methods across countries, we find that the mean expenditure share is about 16.2% for the 2000–2004 window and about 14.6% for the 2005–2009 window, with standard deviations of 5.4% and 3.7%, respectively. [Schafer \(2000\)](#) investigates these shares using older and somewhat more extensive national accounts data and finds that across countries the average expenditure share for household travel is about 11%, with a standard deviation of about 3%. Weighting the household transportation share by 0.6, about the share of expenditure in current US GDP, and adding time costs, we have that the total costs of household travel are between 9% and 11.4% of GDP.

Two further points are made in [Schafer \(2000\)](#). First, for country-level aggregates, per capita travel time and expenditure share are negatively correlated. Second, for Zambia, only 5% of all trips are longer than 10 km, while for the United States, 5% of trips are longer than 50 km. To the extent that these findings are driven by differences in transportation technologies, they suggest that the transferral of developed-country transportation technologies to developing countries is likely to lead to substantial changes in the spatial organization of economic activity.

¹² We note that this estimate is problematic for at least two reasons. First, it assigns the time cost of an average worker to an average traveler, when many travelers are likely to have a lower value of time. Second, it assigns the time cost of an average worker to an average commuter, when wages probably vary systematically with commuting distance. With this said, on the basis of these surveys, a rough guess would be that the aggregate time cost of household travel is somewhere between 3.5% and 8% of the aggregate value of labor in an economy.

Table 20.4 Percentage of household expenditure on all modes of transportation by year and country

	1990–1994	1995–1999	2000–2004	2005–2009	2010–2012
Austria		15	14	16	
Belgium	10	10	13	12	
Bulgaria			23	12	
Canada				15	14
China (rural)	1	3	6	10	10
China (urban)	1	5	9		14
Croatia				13	
Cyprus			23	16	
Czech Republic			26	19	
Denmark		12	11	10	
Estonia			13	16	
Finland		12	14	13	
France	15	13	13	13	
Germany	11	13	12	13	
Greece	12	12	14	12	
Hungary			25	23	
Ireland		14	14	9	
Italy	14	15	15	12	
Latvia			15	19	
Lithuania			17	15	
Luxembourg	12	14	15	16	
Macedonia				16	
Malta			19	25	
Mexico					21
Netherlands	9	9	11	11	
Norway				13	
Poland			19	13	
Portugal	27	23	21	15	
Romania			17	12	
Slovakia			24	15	
Slovenia			26	21	
Spain	14	14	15	12	
Sweden		11	10	12	
Turkey				16	
EU27				13	
United Kingdom					14
United States				18	16

Sources: Statistics Canada (2012b), China Data Online (2011), Eurostat (2009), OECD Stat Extracts (2012), and Bureau of Labour Statistics (2012).

Table 20.5 Carbon dioxide emissions from transportation and shares by mode (2007)

Country	Total CO ₂ equivalents (Mt)	Transport (%)	Rail (%)	Truck (%)	Ship (%)	Air (%)
Canada	561	32	1.32	24.34	1.69	3.14
Mexico	456	35	0.42	32.53	0.57	1.43
United Kingdom	529	29	0.87	25.81	2.38	7.24
United States	5829	33	0.89	27.72	0.95	2.96

Sources: Natural Resources Canada (2012a), North American Transportation Statistics (2012b), and Department for Transport (2012).

20.2.3 External costs

We have so far concerned ourselves with private costs of transportation, time, and private expense. We now turn our attention to two costs of transportation that are rarely priced, carbon emissions and congestion.

Table 20.5 presents total 2007 carbon dioxide equivalent (CO₂e) emissions for the transportation sector for Canada, Mexico, the United States, and the United Kingdom. Total emissions for the United States in 2007 were about 7000 Mt, so the transportation sector accounts for about 30% of US emissions. To the extent that these costs of transportation are not priced, the market allocation of resources to the transportation sector will be, in general, inefficient. With a social cost of about US \$30 per ton CO₂e, the social cost of CO₂e emissions from transportation in the United States is about US \$21 billion per year. This is only about one-tenth of 1% of US GDP. Thus, while greenhouse gas emissions from transportation are important in an absolute sense, they are small relative to the total cost of transportation.

Parry et al. (2007) provides a comprehensive survey of the externalities to automobile use, including local air pollution, global air pollution, traffic congestion, traffic accidents, and other externalities (such as noise and highway maintenance costs). Couture et al. (2012) estimate that a lower bound on the deadweight loss from traffic congestion in the United States is on the order of US \$100 billion per year, although we note that these costs are already reflected in the transportation expenditure data described above.

20.3. THEORETICAL FRAMEWORK

In this section, we outline a multiregion extension of the Helpman (1998) model that follows Redding and Sturm (2008) and Redding (2012). The model incorporates many locations, goods transportation costs within and between locations, and commuting costs within locations. We use the model to show the effects of improvements in transportation infrastructure on the spatial distribution of wages, land rents, the population, and trade within and between locations. Although the model does not capture all of the

theoretical foundations considered in the regional literature and urban literature, it captures many of the standard ingredients, and we use its predictions to structure our review of the empirical evidence below.¹³

20.3.1 Preferences and endowments

The economy consists of a set of locations indexed by n or $i \in N$, where n will typically refer to a consuming region and i to a producing region. To refer to a pairwise quantity, such as a distance or a quantity of trade, we use two subscripts, with the first indicating the location of consumption and the second the location of production. The economy is populated by a mass of representative consumers, $\bar{L}$, who are mobile across locations and are endowed with a single unit of labor that is supplied inelastically with zero disutility. The effective supply of labor for each location i depends on its population (L_i) and commuting technology (b_i), where commuting costs are assumed to take the iceberg form. For each unit of labor residing in location i , only a fraction b_i is available for production, where $0 < b_i < 1$, and the remaining fraction $1 - b_i$ is lost in commuting. While we treat b_i as a primitive of the model here, it could in principle depend on equilibrium population density (e.g., if higher population density increases congestion costs).

Preferences are defined over a consumption index of tradeable varieties, C_n , and consumption of a nontradeable amenity, H_n , which can be interpreted as housing. For simplicity, we treat the stock of housing as a primitive of the model, although it could also in principle depend on equilibrium population density (e.g., if a higher population density increases the supply of housing). The upper level utility function is assumed to be of Cobb–Douglas form¹⁴:

$$U_n = C_n^\mu H_n^{1-\mu}, \quad 0 < \mu < 1. \quad (20.1)$$

The tradeable goods consumption index takes the standard constant elasticity of substitution form:

$$C_n = \left(\sum_{i \in N} M_i c_{ni}^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}},$$

¹³ The model builds on the new economic geography literature synthesized in Fujita et al. (1999). While this literature assumes firm product differentiation and monopolistic competition, the model shares many properties with perfectly competitive models such as the model of Eaton and Kortum (2002) (see Redding, 2012) or the Armington model of product differentiation by location (see Allen and Arkolakis, 2013). The organization of economic activity within countries has recently received renewed attention, as in Cosar and Fajgelbaum (2013) and Ramondo et al. (2012).

¹⁴ For empirical evidence using US data in support of the constant housing expenditure share implied by the Cobb–Douglas functional form, see Davis and Ortalo-Magne (2011).

where σ is the elasticity of substitution between varieties, and we assume that varieties are substitutes ($\sigma > 1$); c_{ni} denotes consumption in country n of a variety produced in country i ; we have used the fact that all varieties produced in location i are consumed in location n in the same amount c_{ni} . Varieties are assumed to be subject to iceberg trade costs. In order for one unit of a variety produced in location i to arrive at location n , a quantity $d_{ni} > 1$ must be shipped, so $d_{ni} - 1$ measures proportional trade costs. The price index dual to the tradeables consumption index C_n is given by

$$P_n = \left(\sum_{i \in N} M_i p_{ni}^{1-\sigma} \right)^{1/(1-\sigma)}, \quad (20.2)$$

where we have used the fact that the measure M_i of varieties produced in location i faces the same elasticity of demand and charges the same equilibrium price $p_{ni} = d_{ni}p_i$ to consumers in location n .

Applying Shephard's lemma to the tradeables price index, we find that the equilibrium demand in location n for a tradeable variety produced in location i is

$$x_{ni} = p_i^{-\sigma} (d_{ni})^{1-\sigma} (\mu v_n L_n) (P_n)^{\sigma-1}, \quad (20.3)$$

where $v_n L_n$ denotes total income, which equals total expenditure, and, with Cobb–Douglas utility, consumers spend a constant share of their income, μ , on tradeables.

With constant expenditure shares and an inelastic supply of the nontradeable amenity, the equilibrium price of this amenity depends solely on the expenditure share, $(1 - \mu)$, total income, $v_n L_n$, and the supply of the nontradeable amenity, $\bar{H}_n$:

$$r_n = \frac{(1 - \mu) v_n L_n}{\bar{H}_n}. \quad (20.4)$$

Total income is the sum of labor income and expenditure on the nontradeable amenity, which is assumed to be redistributed lump-sum to the location's residents:

$$v_n L_n = w_n b_n L_n + (1 - \mu) v_n L_n = \frac{w_n b_n L_n}{\mu}, \quad (20.5)$$

where we have used the fact that only a fraction b_n of the labor in location i is used in production because of commuting costs. Therefore, total labor income equals the wage per effective unit of labor (w_n) times the measure of effective units of labor ($b_n L_n$).

20.3.2 Production technology

There is a fixed cost in terms of labor of producing tradeable varieties ($F > 0$) and a constant variable cost that depends on a location's productivity (A_i). Both the fixed cost and

the variable cost are the same across all varieties produced within a location. The total amount of labor (l_i) required to produce x_i units of a variety in location i is

$$l_i = F + \frac{x_i}{A_i}, \quad (20.6)$$

where we allow productivity (A_i) to vary across locations to capture variation in production fundamentals.

Profit maximization implies that equilibrium prices are a constant markup over marginal cost:

$$p_{ni} = \left(\frac{\sigma}{\sigma - 1} \right) \frac{d_{ni} w_i}{A_i}. \quad (20.7)$$

Combining profit maximization and zero profits, we find the equilibrium output of each tradeable variety equals the following constant:

$$\bar{x} = x_i = \sum_n x_{ni} = A_i F (\sigma - 1). \quad (20.8)$$

Labor market clearing for each location implies that labor demand equals the effective labor supply in that location, which is in turn determined by population mobility. Using the constant equilibrium output of each variety (20.8) and the tradeables production technology (20.6), we can write the labor market clearing condition as follows:

$$b_i L_i = M_i \bar{l}_i = M_i F \sigma, \quad (20.9)$$

where $\bar{l}_i$ denotes the constant equilibrium labor demand for each variety. This relationship pins down the measure of tradeable varieties produced in each location as a function of the location's population, the commuting technology, and the parameters of the model.

20.3.3 Market access and wages

Given demand in all markets and trade costs, the free-on-board price (p_i) charged for a tradeable variety by a firm in each location must be low enough in order to sell the quantity $\bar{x}$ and cover the firm's fixed production costs. We saw above that prices are a constant markup over marginal cost. Therefore, given demand in all markets, the equilibrium wage in location i , w_i , must be sufficiently low in order for a firm to sell $\bar{x}$ and cover its fixed production costs. Using demand (20.3), profit maximization (20.7), and equilibrium output (20.8), we obtain the *tradeables wage equation*:

$$\left(\frac{\sigma}{\sigma - 1} \frac{w_i}{A_i} \right)^\sigma = \frac{1}{\bar{x}} \sum_{n \in N} (w_n b_n L_n) (P_n)^{\sigma-1} (d_{ni})^{1-\sigma}. \quad (20.10)$$

This relationship pins down the maximum wage that a firm in location i can afford to pay given demand in all markets, trade costs, and the production technology. On the right-hand side of the equation, market n demand for tradeables produced in location i depends on the total expenditure on tradeable varieties, $\mu v_n L_n = w_n b_n L_n$, the tradeables price index, P_n , which summarizes the price of competing varieties, and bilateral trade costs, d_{ni} . Total demand for tradeables produced in location i is the weighted sum of the demand in all markets, where the weights are these bilateral trade costs, d_{ni} .

Following Redding and Venables (2004), we define the weighted sum of market demands faced by firms as *firm market access*, fma_i , such that the tradeables wage equation can be written more compactly as

$$w_i = \xi A_i^{\frac{\sigma-1}{\sigma}} [\text{fma}_i]^{1/\sigma}, \quad \text{fma}_i \equiv \sum_{n \in N} (w_n b_n L_n) (P_n)^{\sigma-1} (d_{ni})^{1-\sigma}, \quad (20.11)$$

where $\xi \equiv (F(\sigma-1))^{-1/\sigma}(\sigma-1)/\sigma$ collects together earlier constants. Therefore, wages increase with both productivity A_i and firm market access (fma_i). Investments in transportation infrastructure that reduce the costs of transporting goods (d_{ni}) to markets ($(w_n b_n L_n)(P_n)^{\sigma-1}$) raise market access and wages. Improvements in the commuting technology (b_n) increase the effective supply of labor ($b_n L_n$) and hence total income, which also raises market access and wages.

20.3.4 Labor market equilibrium

With perfect population mobility, workers move across locations to arbitrage away real income differences. Real income in each location depends on per capita income (v_n), the price index for tradeables (P_n), and the price of the nontradeable amenity (r_n). Therefore, population mobility implies

$$V_n = \frac{v_n}{(P_n)^\mu (r_n)^{1-\mu}} = \bar{V}, \quad (20.12)$$

for all locations that are populated in equilibrium, where we have collected the constants $\mu^{-\mu}$ and $(1-\mu)^{-(1-\mu)}$ into the definition of V_n and $\bar{V}$.

The price index (20.2) that enters the above expression for real income depends on consumers' access to tradeable varieties, as captured by the measure of varieties and their free-on-board prices in each location i , together with the trade costs of shipping the varieties from location i to location n . We summarize consumers' access to tradeables using the concept of *consumer market access*, cma_n :

$$P_n = (\text{cma}_n)^{1/(1-\sigma)}, \quad \text{cma}_n \equiv \sum_{i \in N} M_i (p_i d_{ni})^{1-\sigma}. \quad (20.13)$$

Substituting for v_n , P_n , and r_n , we can rewrite the labor mobility condition (20.12) to yield an expression linking the equilibrium population of a location (L_n) to its productivity

(A_n) , its commuting technology (b_n) , the supply of the nontraded amenity (H_n) , and the two endogenous measures of market access introduced above (one for firms (fma_n) and one for consumers (cma_n)):

$$L_n = \chi b_n^{\frac{\mu}{1-\mu}} A_n^{\frac{\mu(\sigma-1)}{\sigma(1-\mu)}} \bar{H}_n (fma_n)^{\frac{\mu}{\sigma(1-\mu)}} (cma_n)^{\frac{\mu}{(1-\mu)(\sigma-1)}}, \quad (20.14)$$

where $\chi = \bar{V}^{-1/(1-\mu)} \xi^{\mu/(1-\mu)} \mu^{-\mu/(1-\mu)} (1-\mu)^{-1}$ is a function of the common real income $\bar{V}$.

Therefore, the equilibrium population (L_n) increases with the quality of the commuting technology (b_n) , the productivity of the final goods production technology (A_n) , and the supply of the nontraded amenity $(\bar{H}_n)$. Investments in transportation infrastructure that reduce the costs of transporting goods (d_{ni}) raise both firm market access and consumer market access (fma_n) and (cma_n) and hence increase the equilibrium population. Improvements in the commuting technology (b_n) also have positive indirect effects on the equilibrium population through higher firm and consumer market access.

From land market clearing (20.4) and total labor income (20.5), land prices can be written in terms of wages and the total population:

$$r_n = \frac{(1-\mu) w_n b_n L_n}{\mu \bar{H}_n}. \quad (20.15)$$

Therefore, higher firm market access (fma_n) raises land prices through both higher wages (from (20.10)) and higher population (from (20.14)), while higher consumer market access (cma_n) raises land prices through a higher population alone (from (20.14)). Reductions in the cost of transporting goods (d_{ni}) raise land prices through both firm market access and consumer market access. Improvements in commuting technology (b_n) raise land prices directly and also indirectly through higher wages and a higher population.

20.3.5 Trade flows

Using constant elasticity of substitution demand, we can express the share of location n 's expenditure on varieties produced in location i as

$$\pi_{ni} = \frac{M_i p_{ni}^{1-\sigma}}{\sum_{k \in N} M_k p_{nk}^{1-\sigma}}, \quad (20.16)$$

which, with use of the equilibrium pricing rule (20.7) and the labor market clearing condition for each location (20.9), can be written as

$$\pi_{ni} = \frac{b_i L_i (d_{ni} w_i)^{1-\sigma} (A_i)^{\sigma-1}}{\sum_{k \in N} b_k L_k (d_{nk} w_k)^{1-\sigma} (A_k)^{\sigma-1}}. \quad (20.17)$$

This expression for bilateral trade shares (π_{ni}) corresponds to a “gravity equation,” in which bilateral trade between exporter i and importer n depends on both “bilateral resistance” (i.e., the bilateral goods of trading goods between exporter i and importer n (d_{ni}) in the numerator) and “multilateral resistance” (i.e., the bilateral costs for importer n of sourcing goods from all exporters k (d_{nk}) in the denominator). In this gravity equation specification, bilateral trade depends on characteristics of the exporter i (e.g., the exporter’s wage w_i in the numerator), bilateral trade costs (d_{ni}), and characteristics of the importer n (i.e., the importer’s access to all sources of supply in the denominator).¹⁵

Taking the ratio of these expenditure shares, we find the value of trade between locations (X_{ni}) relative to trade within locations (X_{nn}) is

$$\frac{X_{ni}}{X_{nn}} = \frac{\pi_{ni}}{\pi_{nn}} = \frac{b_i L_i (d_{ni} w_i)^{1-\sigma} (A_i)^{\sigma-1}}{b_n L_n (d_{nn} w_n)^{1-\sigma} (A_n)^{\sigma-1}}. \quad (20.18)$$

Therefore, transportation infrastructure improvements that reduce the cost of transporting goods within locations (d_{nn}) by the same proportion as they reduce the cost of transporting goods between locations (d_{ni}) leave the ratio of trade between locations to trade within locations unchanged. One potential example is building roads within cities that make it easier for goods to circulate within the city and to leave the city to connect with long-distance highways. Transportation cost improvements that reduce commuting costs for all locations (increase b_n and b_i) also leave the ratio of trade between locations to trade within locations unchanged.

In this model with a single differentiated sector, all trade takes the form of intraindustry trade, and transportation infrastructure improvements affect the volume of this intraindustry trade. More generally, in a setting with multiple differentiated sectors that differ in terms of the magnitude of trade costs (e.g., high value to weight versus low value to weight sectors), transportation infrastructure improvements also affect the pattern of interindustry trade and the composition of employment and production across sectors within locations.

20.3.6 Welfare

We now show how the structure of the model can be used to derive an expression for the welfare effects of transportation infrastructure improvements in terms of observables. Using the trade share (20.16), we can rewrite the price index (20.2) in terms of each location’s trade share with itself and other parameters:

¹⁵ For an insightful review of the gravity equation in the international trade literature, see [Head and Mayer \(2013\)](#).

$$P_n = \frac{\sigma}{\sigma - 1} \left(\frac{b_n L_n}{\sigma F \pi_{nn}} \right)^{\frac{1}{1-\sigma}} \frac{d_{nn} w_n}{A_n}. \quad (20.19)$$

From this expression for the price index and land market clearing (20.15), the population mobility condition (20.12) implies that the equilibrium population for each location can be written as

$$L_n = \left(\frac{\left(\frac{1}{\sigma F \pi_{nn}} \right)^{\frac{\mu}{\sigma-1}} \bar{H}_n^{1-\mu} b_n^{\frac{\mu\sigma}{\sigma-1}} A_n^\mu}{\mu \left(\frac{1-\mu}{\mu} \right)^{1-\mu} \left(\frac{\sigma}{\sigma-1} \right)^\mu \bar{V} d_{nn}^\mu} \right)^{\frac{\sigma-1}{\sigma(1-\mu)-1}}, \quad (20.20)$$

where terms in wages (w_n) have canceled and labor market clearing for the economy as a whole implies

$$\sum_{n \in N} L_n = \bar{L}. \quad (20.21)$$

This expression for the equilibrium population (20.20) has an intuitive interpretation. The population of each location n decreases with its domestic trade share (π_{nn}), since locations with low domestic trade shares have low consumption goods price indices, which increases their attractiveness to residents. The population of each location increases with the efficiency of its commuting technology (b_n), its productivity in production (A_n), its supply of housing ($\bar{H}_n$), and its transportation technology (inversely related to d_{nn}). The common level of utility across all locations ($\bar{V}$) is endogenous and is determined by the requirement that the labor market clears for the economy as a whole.

Rearranging the population mobility condition (20.20), we can write the real income in each location in terms of its population, trade share with itself, and other parameters.

$$V_n = \frac{\left(\frac{1}{\sigma F \pi_{nn}} \right)^{\frac{\mu}{\sigma-1}} L_n^{\frac{\sigma(1-\mu)-1}{\sigma-1}} \bar{H}_n^{1-\mu} b_n^{\frac{\mu\sigma}{\sigma-1}} A_n^\mu}{\mu \left(\frac{1-\mu}{\mu} \right)^{1-\mu} \left(\frac{\sigma}{\sigma-1} \right)^\mu d_{nn}^\mu} = \bar{V}. \quad (20.22)$$

A key implication of this expression for real income is that the change in each location's trade share with itself and the change in its population are sufficient statistics for the welfare effects of improvements in transportation technology that reduce the costs of trading goods (see Redding, 2012):

$$\frac{V_n^1}{V_n^0} = \left(\frac{\pi_{nn}^0}{\pi_{nn}^1} \right)^{\frac{\mu}{\sigma-1}} \left(\frac{L_n^0}{L_n^1} \right)^{\frac{\sigma(1-\mu)-1}{\sigma-1}} = \frac{\bar{V}^1}{\bar{V}^0}, \quad (20.23)$$

where the superscripts 0 and 1 denote the value of variables before and after the improvement in transportation technology, respectively.

Similar sufficient statistics apply for the welfare effects of improvements in transportation technology that reduce commuting costs, although these welfare effects also depend directly on the change in commuting costs (through the resulting increase in the effective supply of labor):

$$\frac{V_n^1}{V_n^0} = \left(\frac{b_n^1}{b_n^0} \right)^{\frac{\mu\sigma}{\sigma-1}} \left(\frac{\pi_{nn}^0}{\pi_{nn}^1} \right)^{\frac{\mu}{\sigma-1}} \left(\frac{L_n^0}{L_n^1} \right)^{\left(\frac{\sigma(1-\mu)-1}{\sigma-1} \right)} = \frac{\bar{V}^1}{\bar{V}^0}. \quad (20.24)$$

While these improvements in transportation infrastructure have uneven effects on wages, land prices, and the population, the mobility of workers across locations ensures that they have the same effect on welfare across all populated locations.

To understand the relationship between changes in domestic trade shares and the welfare change from improvements in transportation technology that reduce goods trade costs, consider the extreme case where the transportation improvement allows goods trade between two previously autarkic locations. For locations closed to goods trade, domestic trade shares must equal 1. Once locations open to trade, they can specialize to exploit gains from trade with other locations, and domestic trade shares fall below 1. This fall in the domestic trade shares reflects the increase in specialization and is directly related to increases in real income, our measure of welfare.

To understand the relationship between changes in population and the changes in welfare following improvements in transportation technology that reduce goods trade costs, first note that labor mobility requires real wage equalization across populated locations. Therefore, if goods trade is opened between locations, and some locations (e.g., coastal regions) benefit more than other locations (e.g., interior regions) at the initial labor allocation, workers must relocate to arbitrage away real wage differences. Those locations that experience larger welfare gains from trade at the initial labor allocation will experience population inflows, which increases the demand for the immobile factor land and causes land prices to rise. In contrast, those locations that experience smaller welfare gains from trade at the initial labor allocation will experience population outflows, which decreases the demand for land and reduces land prices. This population reallocation continues until real wages are again equalized across all populated locations. Hence, these population changes also need to be taken into account in computing the welfare effects of the improvement in transportation technology.

Therefore, together, the change in a location's domestic trade share and the change in its population are sufficient statistics for the effects of a transportation improvement that reduces the costs of trading goods (d_{nn}). A transportation improvement that reduces the commuting costs for a region (b_n) also directly increases the supply of labor for that region, which is taken into account in the welfare formula.

20.3.7 General equilibrium

The general equilibrium of the model can be represented by the share of workers in each location ($\lambda_n = L_n/\bar{L}$), the share of each location's expenditure on goods produced by other locations (π_{ni}), and the wage in each location (w_n). Using labor income (20.5), the trade share (20.16), population mobility (20.20), and labor market clearing (20.21), the equilibrium triple $\{\lambda_n, \pi_{ni}, w_n\}$ solves the following system of equations for all $i, n \in N$ (see Redding, 2012):

$$w_i b_i \lambda_i = \sum_{n \in N} \pi_{ni} w_n b_n \lambda_n, \quad (20.25)$$

$$\pi_{ni} = \frac{b_i \lambda_i (d_{ni} w_i / A_i)^{1-\sigma}}{\sum_{k \in N} b_k \lambda_k (d_{nk} w_k / A_k)^{1-\sigma}}, \quad (20.26)$$

$$\lambda_n = \frac{\left[\bar{H}_n^{1-\mu} \left(\frac{1}{\pi_{nn}} \right)^{\frac{\mu}{\sigma-1}} b_n^{\frac{\mu\sigma}{\sigma-1}} A_n^\mu d_{nn}^{-\mu} \right]^{\frac{\sigma-1}{\sigma(1-\mu)-1}}}{\sum_{k \in N} \left[\bar{H}_k^{1-\mu} \left(\frac{1}{\pi_{kk}} \right)^{\frac{\mu}{\sigma-1}} b_k^{\frac{\mu\sigma}{\sigma-1}} A_k^\mu d_{kk}^{-\mu} \right]^{\frac{\sigma-1}{\sigma(1-\mu)-1}}}. \quad (20.27)$$

The assumption that $\sigma(1 - \mu) > 1$ corresponds to the “no black hole” condition in Krugman (1991) and Helpman (1998). For parameter values satisfying this inequality, the model's agglomeration forces from love of variety, increasing returns to scale, and transportation costs (which are inversely related to σ) are not too strong relative to its congestion forces from an inelastic supply of land (captured by $1 - \mu$). As a result, each location's real income monotonically decreases with its population, which ensures the existence of a unique stable nondegenerate distribution of the population across locations.

While the existence of a unique equilibrium ensures that the model remains tractable and amenable to counterfactual analysis, often the rationale for transportation investments is cast in terms of shifting the distribution of economic activity between multiple equilibria. To the extent that such multiple equilibria exist, their analysis requires either consideration of the range of the parameter space for which the model has multiple equilibria or the use of a richer theoretical framework.¹⁶

¹⁶ A body of empirical literature has examined whether large and temporary shocks have permanent effects on the location of economic activity and has interpreted these permanent effects as either evidence of multiple equilibria or path dependence more broadly. See, for example, Bleakley and Lin (2012), Davis and Weinstein (2002), Maystadt and Duranton (2014), Redding et al. (2011), and Sarvimäki et al. (2010).

20.3.8 Counterfactuals

The system of equations for general equilibrium (20.25)–(20.27) can be used to undertake model-based counterfactuals in an extension of the trade-based approach of Dekle et al. (2007) to incorporate factor mobility across locations. The system of equations for general equilibrium must hold both before and after any counterfactual change in, for example, transportation infrastructure. Denote the value of variables in the counterfactual equilibrium with a prime (x') and the relative value of variables in the counterfactual and initial equilibria by a hat ($\hat{x} = x' / x$). Using this notation, we can rewrite the system of equations for the counterfactual equilibrium (20.25)–(20.27) as follows:

$$\hat{w}_i \hat{b}_i \hat{\lambda}_i Y_i = \sum_{n \in N} \hat{\pi}_{ni} \pi_{ni} \hat{w}_n \hat{b}_n \hat{\lambda}_n Y_n, \quad (20.28)$$

$$\hat{\pi}_{ni} \pi_{ni} = \frac{\pi_{ni} \hat{\lambda}_i \hat{b}_i (\hat{d}_{ni} \hat{w}_i / \hat{A}_i)^{1-\sigma}}{\sum_{k \in N} \pi_{nk} \hat{\lambda}_k \hat{b}_k (\hat{d}_{nk} \hat{w}_k / \hat{A}_k)^{1-\sigma}}, \quad (20.29)$$

$$\hat{\lambda}_n \lambda_n = \frac{\lambda_n \left(\hat{H}^{1-\mu} \hat{\pi}_{nn}^{\frac{\mu}{\sigma-1}} \hat{b}_n^{\frac{\mu\sigma}{\sigma-1}} \hat{A}_n^\mu \hat{d}_{nn}^{-\mu} \right)^{\frac{\sigma-1}{\sigma(1-\mu)-1}}}{\sum_{k \in N} \lambda_k \left(\hat{H}_k^{1-\mu} \hat{\pi}_{kk}^{\frac{\mu}{\sigma-1}} \hat{b}_k^{\frac{\mu\sigma}{\sigma-1}} \hat{A}_k^\mu \hat{d}_{kk}^{-\mu} \right)^{\frac{\sigma-1}{\sigma(1-\mu)-1}}}, \quad (20.30)$$

where $Y_i = w_i b_i L_i$ denotes labor income in the initial equilibrium.

Given an exogenous change in transportation infrastructure that affects the costs of trading goods ($\hat{d}_{ni}$) or the costs of commuting ($\hat{b}_n$), this system of equations (20.28)–(20.30) can be solved for the counterfactual changes in wages ($\hat{w}_n$), population shares ($\hat{\lambda}_n$), and trade shares ($\hat{\pi}_{ni}$). Implementing these counterfactuals requires only observed values of GDP, trade shares, and population shares $\{Y_n, \pi_{ni}, \lambda_n\}$ for all locations $i, n \in N$ in the initial equilibrium. For parameter values for which the model has a unique stable equilibrium ($\sigma(1 - \mu) > 1$), these counterfactuals yield determinate predictions for the impact of the change in transportation costs. From the welfare analysis above, the changes in each location's population and its domestic trade share provide sufficient statistics for the welfare effect of transportation improvements that affect the costs of trading goods ($\hat{d}_{ni}$). In contrast, transportation improvements that affect the costs of commuting ($\hat{b}_n$) also have direct effects on welfare in addition to their effects through the population and domestic trade shares. With perfect population mobility, these welfare effects must be the same across all populated locations.

20.4. REDUCED-FORM ECONOMETRIC FRAMEWORK

20.4.1 A simple taxonomy

We survey the recent empirical literature investigating the effects of infrastructure on the geographic distribution of economic activity. The preponderance of this literature can be described with a remarkably simple taxonomy.

Let t index time periods, and, preserving the notation from above, let n and $i \in N$ index a set of geographic locations, typically cities or counties. Let L_{it} denote an outcome of interest for location i at time t : employment, population, rent, or centralization. Let x_{it} be a vector of location- and time-specific covariates, and finally, let b_{it} and d_{it} denote the transportation variables of interest. In particular, consistent with the notation in our theoretical model, let b_{it} denote a measure of transportation infrastructure that is internal to unit i , and let d_{it} denote a measure of transportation infrastructure external to unit i . For example, b_{it} could count radial highways within a metropolitan area, while d_{it} could indicate whether a rural county is connected to a highway network.

With this notation in place, define the “intracity regression” as

$$L_{it} = C_0 + C_1 b_{it} + C_2 x_{it} + \delta_i + \theta_t + \epsilon_{it}, \quad (20.31)$$

where δ_i denotes location-specific time-invariant unobservables, θ_t denotes a common time effect for all locations, and ϵ_{it} denotes the time-varying location-specific residual. The coefficient of interest is C_1 , which measures the effect of within-city infrastructure on the city-level outcome.¹⁷

Similarly, define the “intercity regression” as

$$L_{it} = C_0 + C_1 d_{it} + C_2 x_{it} + \delta_i + \theta_t + \epsilon_{it}, \quad (20.32)$$

which differs from the intracity regression only in that the explanatory variable of interest describes transportation costs between unit i and other units, rather than within-city infrastructure.

These equations require some discussion before we turn to a description of the results. First, both estimating equations are natural reduced-form versions of Equation (20.14) or, if the outcome of interest is land rent, Equation (20.15). Thus, they are broadly consistent with the theoretical framework described earlier. Second, comparing the regression equations with their theoretical counterparts immediately suggests four inference problems that estimations of the intracity and intercity regressions should confront.

First, equilibrium employment or land rent depends on the location-specific productivity, A_n . This will generally be unobserved and thus will be reflected in the error terms of our regression equations. It is natural to expect that intracity and intercity infrastructure will depend on location-specific productivity and, hence, be endogenous in the two

¹⁷ Moses (1958) and Moses and Williamson (1963) are pioneering studies on the role of automobiles and highways in reorganizing the distributions of population and economic activity within metropolitan areas.

regression equations. Second, equilibrium employment or land rent depends on the level of a location-specific amenity, H_n . In our model, this reflects a supply of housing, but in reality, it may also reflect unobserved location characteristics that augment or reduce the welfare of residents at a location. We might also be concerned that such amenities, to the extent that they are unobserved, affect infrastructure allocation and give rise to an endogeneity problem. More generally, the intercity and intracity regressions do not by themselves distinguish between the demand for and supply of transportation.

Third, Equations (20.14) and (20.15) involve expressions for market access not present explicitly in the estimating equations. To the extent that market access depends on transportation costs between cities, the treatment of market access in these estimations deserves careful attention. Fourth, to the extent that there are general equilibrium effects of transportation infrastructure on all locations, these are not captured by C_1 . Instead, they are captured in the time effects θ_t and cannot be separated from other time-varying factors that are common to all locations without further assumptions. More generally, in general equilibrium, transportation investments between a pair of regions i and j can have effects on third regions k , which are not captured by the transportation variables for regions i and j .

20.4.2 Identification of causal effects

As discussed above, perhaps the biggest empirical challenge in estimating the intercity and intracity regressions is constructing the appropriate counterfactual for the absence of the transportation improvement. In particular, ordinary least squares (OLS) regressions comparing treated and untreated locations are unlikely to consistently estimate the causal effect of the transportation improvement, because the selection of locations into the treatment group is nonrandom. The main empirical approach to addressing this challenge has been to develop instruments for the assignment of transportation improvements that plausibly satisfy the exclusion restriction of affecting the economic outcome of interest only through the transportation improvement.¹⁸ More formally, this approach to identifying the causal effects posits an additional first-stage regression that determines the assignment of transportation infrastructure:

$$\mathbb{I}_{it} = D_0 + D_1 x_{it} + D_2 z_{it} + \eta_i + \gamma_t + u_{it}, \quad (20.33)$$

where $\mathbb{I}_{it} \in \{b_{it}, d_{it}\}$ is the transportation variable of interest (depending on whether the specification is intracity or intercity), x_{it} are the location- and time-varying controls from the second-stage regression ((20.31) or (20.32)), η_i are location-specific time-invariant

¹⁸ While the program evaluation literature suggests other complementary approaches, such as conducting randomized experiments with transportation improvements or the use of matching estimators, these have been less widely applied in this empirical literature.

unobservables, γ_t are time indicators, u_{it} is a time-varying location-specific residual, and z_{it} are the instruments or excluded exogenous variables.

Combining the second-stage equation ((20.31) or (20.32)) with the first-stage equation (20.33), we can estimate the impact of transportation infrastructure on the economic outcomes of interest (C_1) using two-stage least squares. Credible identification of the causal impact of transportation infrastructure requires that two conditions are satisfied: (1) the instruments have power in the first-stage regression ($D_2 \neq 0$) and (2) the instruments satisfy the exclusion restriction of affecting the economic outcomes of interest only through transportation infrastructure conditional on the controls x_{it} —that is, $\text{cov}(\epsilon_{it}, u_{it}) = 0$.

The existing literature has followed three main instrumental variables strategies. The first, the *planned route instrumental variable approach*, is an instrumental variables strategy which relies on planning maps and documents as a source of quasi-random variation in the observed infrastructure. The second, the *historical route instrumental variable approach*, relies on very old transportation routes as a source of quasi-random variation in observed infrastructure. The third, the *inconsequential place approach*, relies on choosing a sample that is inconsequential in the sense that unobservable attributes do not affect the placement of infrastructure. The plausibility of these identification strategies depends sensitively on the details of their implementation and is sometimes contentious. With this said, we briefly describe these identification strategies and the rationale for their use. We avoid discussion of the validity of these strategies in particular contexts. Broadly, the strategies we describe are the best approaches currently available for estimating the causal effects of transportation infrastructure on the organization of economic activity.

20.4.2.1 Planned route instrumental variable approach

Baum-Snow (2007) pioneers the planned route instrumental variable approach by using a circa 1947 plan for the interstate highway network as a source of quasi-random variation in the way the actual network was developed. In the specific context of Baum-Snow (2007), this means counting the number of planned radial highways entering a metropolitan area and using this variable to predict the actual number of interstate highway rays. Since the network plan was developed under a mandate to serve military purposes, the validity of this instrument hinges on the extent to which military purposes are orthogonal to the needs of postwar commuters. Several other empirical investigations into the effects of the US road and highway network exploit instruments based on the 1947 highway plan, while Hsu and Zhang (2012) develop a similar instrument for Japan. Michaels et al. (2012) use an even earlier plan of the US highway network, the “Pershing plan,” as a source of quasi-random variation in the US highway network. Although Donaldson (2015) stops short of using hypothetical planned networks as instruments for realized networks, he does compare the development of districts without railroads and without planned railroads with those without railroads but with planned railroads.

That these sets of districts develop in the same way suggests that the planning process did not pick out districts on the basis of different unobservable characteristics.

20.4.2.2 Historical route instrumental variable approach

[Duranton and Turner \(2012\)](#) develop the historical route instrumental variable approach. In regressions predicting metropolitan statistical area (MSA)-level economic outcomes they rely on maps of historical transportation networks, the US railroad network circa 1898, and the routes of major expeditions of exploration of the United States between 1535 and 1850 as sources of quasi-random variation in the US interstate highway network at the end of the twentieth century. The validity of these instruments requires that, conditional on the controls, factors that do not directly affect economic activity in US metropolitan areas at the end of the twentieth century determine the configuration of these historical networks. A series of articles ([Duranton and Turner, 2011, 2012](#); [Duranton et al., 2014](#)) use the two historical route instruments and the 1947 highway plan as sources of quasi-random variation in regressions predicting metropolitan total vehicle kilometers traveled, changes in metropolitan employment, and trade flows between cities as functions of the interstate highway network.

One distinctive feature of [Duranton and Turner \(2011, 2012\)](#) and [Duranton et al. \(2014\)](#) is the use of multiple instruments based on different sources of variation. With more instruments than endogenous variables, the specification can be estimated with either all or subsets of the instruments, and overidentification tests can be used as a check on the identifying assumptions. Conditional on one of the instruments being valid, these overidentification tests check the validity of the other instruments. Given that the instruments exploit quite different sources of variation in the data, if a specification passes the overidentification test, this implies that either all of the instruments are valid or an improbable relationship exists between the instruments and the errors of the first-stage and second-stage regressions.

Several other authors develop historical transportation networks as a source of quasi-random variation in modern transportation networks in other regions. [Baum-Snow et al. \(2012\)](#) rely on Chinese road and rail networks from 1962 as a source of quasi-random variation in road and rail networks after 2000. [Garcia-Lopez et al. \(2013\)](#) use eighteenth century postal routes and Roman roads for Spain. [Hsu and Zhang \(2012\)](#) rely on historical Japanese railroad networks. [Martincus et al. \(2012\)](#) use the Inca roads for Peru. [Duranton and Turner \(2012\)](#) provide a more detailed discussion of the validity of these instruments.

20.4.2.3 Inconsequential units approach

To estimate the intercity regression, researchers often rely on the inconsequential units approach to identification, sometimes in conjunction with one or both of the instrumental variables strategies described above. If we consider economically small units lying

between large cities, then we expect that intercity links will traverse these units only when they lie along a convenient route between the two large cities. That is, we expect that the unobserved characteristics of units between large cities are inconsequential to the choice of route, and therefore that the connection status of these units will not depend on the extent to which these units are affected by the road. Chandra and Thompson (2000) pioneer this strategy in their analysis of the effect of access to the interstate highway system on rural counties in the United States. By restricting attention to rural highways, they hope to restrict attention to counties that received interstate highways “accidentally,” by virtue of lying between larger cities. While it is difficult to assess the validity of this approach, some of the regressions reported in Michaels (2008) are quite similar to those in Chandra and Thompson (2000) but rely on the 1947 planned highway network for identification. That the two methods arrive at similar estimates is reassuring. Banerjee et al. (2012) also use the inconsequential units strategy in their analysis of the effects of Chinese transportation networks. In particular, they construct a hypothetical transportation network connecting historical treaty ports to major interior trading centers. Counties near these predicted networks are there accidentally in the same sense that rural counties may be accidentally near interstate highways in the United States. Similarly, and also for China, Faber (2015) constructs a hypothetical least-cost network connecting major Chinese cities and examines the impact of proximity to this network on outcomes in nearby rural counties.

These three econometric responses to the probable endogeneity of transportation infrastructure are widely used. Other approaches to this problem typically exploit natural experiments that, while they may provide credible quasi-random variation in infrastructure, are not easily extended to other applications.

20.4.3 Distinguishing growth from reorganization

As Fogel (1964) observes in his classic analysis of the role of railroad construction in the economic development of the nineteenth century United States, an assessment of the economic impacts of transportation infrastructure depends fundamentally on whether changes in transportation costs affect the amount of economic activity or reorganize existing economic activity. For example, the welfare implications of a road or light rail line that attracts preexisting firms are quite different from those of one that leads to the creation of new firms. Importantly, this issue is distinct from the endogeneity problem discussed above. The problem of endogeneity follows from nonrandom assignment of transportation infrastructure to “treated” observations. The problem of distinguishing between growth and reorganization persists even when transportation is assigned to observations at random. Even in the case in which a region experiences an exogenous change in transportation infrastructure, the observed effects on economic activity in the region can reflect either reorganization or growth. This same issue of distinguishing

growth and reorganization appears in the literature evaluating place-based policies, as discussed in [Neumark and Simpson \(2014\)](#) in this volume.¹⁹

Figure 20.6 illustrates a simple hypothetical dataset with the same structure as that typically used to estimate the intercity and intracity estimating equations. Figure 20.6 describes a sample consisting of three regions: a region that is “treated” in some way that affects transportation costs in this region—for example, a new road; an untreated region which is typically near the treated region but is not subject to a change in transportation infrastructure; and everywhere else. The outcome variable of interest is y and the new road creates a units of this outcome in the treated region and displaces d units from the untreated region to the treated region.

Fundamentally, the intercity and intracity regressions estimate the effect of treatment on the difference between treated regions and untreated comparison regions. As the figure makes clear, the difference in the outcome between treated and untreated regions is $2d + a$, the compound effect of reorganization and growth. At its core, the problem of distinguishing between reorganization and growth requires us to identify two quantities. Without further assumptions, these two quantities cannot be separately identified if we estimate only a single equation, regardless of whether it is the intercity or intracity estimating equation. To identify both the growth and the reorganization effect, we must estimate two linearly independent equations.

In the context of the sample described in Figure 20.6, these two equations could involve a comparison of any two of the three possible pairs of regions—that is, treated

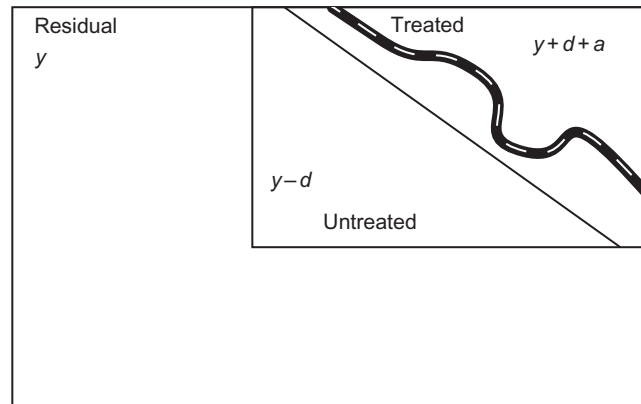


Figure 20.6 A simple hypothetical sample.

¹⁹ For approaches to distinguishing growth and reorganization in this literature on place-based policies, see [Criscuolo et al. \(2012\)](#) and [Mayer et al. \(2013\)](#).

and untreated, untreated and residual, and treated and residual. Alternatively, with panel data, one could estimate the change in the treated region following the change in transportation costs and also the change in the untreated region following the change in the treated region. While the literature has carefully addressed the possibility that transportation costs and infrastructure are not assigned to regions at random, few authors conduct estimations allowing the separate identification of growth and reorganization.

While [Figure 20.6](#) suggests simple methods for distinguishing between growth and reorganization, this reflects implicit simplifying assumptions. In particular, the new road in the treated district does not lead to migration of economic activity from the residual region to the untreated or the treated region and does not cause growth in the untreated or residual region. If we allow these effects, then the effect of a new road in the treated region is characterized by six parameters rather than two. Identifying all of these parameters will generally require estimating six linearly independent equations and will not generally be possible with cross-sectional data. In the context of “real data,” with a more complex geography and many regions subject to treatment, distinguishing between growth and reorganization requires *a priori* restrictions on the nature of these effects.

The literature has, as yet, devoted little attention to what these identifying assumptions should be. As suggested by [Figure 20.6](#), this problem can be resolved with transparent but ad hoc assumptions. Alternatively, the theoretical model described in [Section 20.3](#) provides a theoretically founded basis for distinguishing between growth and reorganization which derives from the iceberg structure of transportation costs and assumptions about demand and production. Importantly, if the new road in the treated region affects the level of economic activity in all three regions, then no cross-sectional estimate can recover this effect. This requires time series data or cross-sectional data describing “replications” of [Figure 20.6](#). More generally, for a penetration road or single transport project, it may be possible to construct plausible definitions of treated, untreated, and residual regions, as in [Figure 20.6](#). However, for an evaluation of a national highway system, there may be no plausible residual regions, in which case we are necessarily in a general equilibrium world.

20.5. REDUCED-FORM EMPIRICAL RESULTS

20.5.1 Intracity infrastructure and the geographic organization of economic activity

20.5.1.1 Infrastructure and decentralization

[Baum-Snow \(2007\)](#) partitions a sample of US metropolitan areas into an “old central business district,” the central business district circa 1950, and the residual suburbs. He then estimates a version of the intracity regression, Equation (20.31), in first differences, where the unit of observation is a US MSA, the measure of infrastructure is the count of radial interstate highways, and the instrument is a measure of rays based on the 1947

highway plan discussed above. He finds that each radial segment of the interstate highway network causes about a 9% decrease in the central city population. Since one standard deviation in the number of rays in an MSA is 1.5, this means that a one standard deviation increase in the number of rays causes an about 14% decrease in the central city population. To get a sense of the magnitude of this effect, the US population grew by 64% during his study period, the MSA population grew by 72%, and the constant-boundary central city population declined by 17%. Thus, the interstate highway system can account for almost the entire decline in old central city population densities. Note that, since [Baum-Snow \(2007\)](#) estimates the share of the population in the treated area, he avoids the problem of distinguishing between growth and reorganization. The share of the population in the central city reflects changes in the level of the central city and the suburb and migration between the two.

This result has been extended to two other contexts. [Baum-Snow et al. \(2012\)](#) conduct essentially the same regression using data describing Chinese prefectures between 1990 and 2010. They first partition each prefecture into the constant-boundary administrative central city and the residual prefecture and then examine the effect of several measures of infrastructure on the decentralization of the population and employment. They rely on historical routes (from 1962) as a source of quasi-random variation in city-level infrastructure. They find that each major highway ray causes an about 5% decrease in the central city population. No other measure of infrastructure—kilometers of highways, ring road capacity, kilometers of railroads, ring rail capacity, or radial rail capacity—has a measurable effect on the organization of the population in Chinese prefectures. [Baum-Snow et al. \(2012\)](#) also examine the effect of infrastructure on the organization of production. They find that radial railroads and highway ring capacity both have dramatic effects on the organization of production. In particular, each radial railroad causes about 26% of central city manufacturing to migrate to the periphery, while ring roads also have a dramatic effect. This effect varies by industry. Industries with relatively low weight-to-value ratios are more affected. None of the other infrastructure measures they investigate affect the organization of production.

Finally, [Garcia-Lopez et al. \(2013\)](#) consider the effect of limited-access highways on the organization of the population in Spanish cities between 1991 and 2011. Their unit of observation is 1 of 123 Spanish metropolitan regions. They conduct a version of the intra-city regression in first differences to explain the change in the central city population between 1991 and 2011 as a function of changes in the highway network over the same period. They rely on three historical road networks as an instrument for changes in the modern network: the Roman road network; a network of postal roads, circa 1760; and a network of nineteenth century main roads. They find that each radial highway causes an about 5% decrease in the central city population, and that kilometers of central city or suburban highways have no measurable effect. Using a similar instrumentation strategy, [Ángel García-López \(2012\)](#) examines the impact of transportation improvements on the

location of the population within the city of Barcelona. Consistent with some of the findings discussed above, improvements to the highway and railroad systems are found to foster population growth in suburban areas, whereas the expansion of the transit system is found to affect the location of the population inside the central business district.²⁰

While the decentralization articles mentioned above investigate the effect on central cities of infrastructure improvements which reduce the cost of accessing peripheral land, [Ahlfeldt et al. \(2014\)](#) consider the effect of changes in transportation cost between two adjacent parts of the same central city. Specifically, [Ahlfeldt et al. \(2014\)](#) consider the effect of the construction and destruction of the Berlin Wall, which separated West Berlin from the historical central business district. They examine population, employment, and land values in 1936, before the partition of the city, in 1986, shortly before reunification, and in 2006, 15 years after reunification. That is, when the cost of commuting from the West to the East was low, prohibitively high, and low again.

Methodologically, the analysis by [Ahlfeldt et al. \(2014\)](#) differs dramatically from the analyses in the centralization articles mentioned above. Their sample consists of approximately 16,000 “statistical blocks” comprising metropolitan Berlin, each with a population of about 250 people in 2005. Loosely, for each block, [Ahlfeldt et al. \(2014\)](#) record the location, population, land rent, and employment in the 3 years of their study. They use these data to estimate a first-differences variant of the intercity regression (Equation 20.32). The reduced-form results in [Ahlfeldt et al. \(2014\)](#) show that the construction of the Berlin Wall caused the central business district to migrate so that it was more nearly central in the territory of West Berlin, and that the removal of the Berlin Wall approximately reversed this process. The identifying assumption underlying this natural experiment is that change in access to economic activity following from division and reunification is uncorrelated with other changes in the way the city was organized, except through its effect on access to economic activity. In addition to these reduced-form results, [Ahlfeldt et al. \(2014\)](#) also conduct structural estimations, which we discuss later.

20.5.1.2 Infrastructure and miscellaneous city-level outcomes

Beyond the literature investigating infrastructure and decentralization, a series of articles by Duranton and Turner investigate the relationship between roads and employment growth, intercity trade, and driving.

[Duranton and Turner \(2012\)](#) examine employment growth in US MSAs between 1984 and 2004. Their principal regression is a variant of the intracity regression for which the outcome is employment growth between 1984 and 2004, and their measure of transportation is kilometers of interstate highways within city boundaries. They rely on the

²⁰ One issue that has received relatively little attention in the intracity literature is the role of transportation infrastructure in segregating cities and leaving some neighborhoods “on the wrong side of the tracks.”

1947 highway plan, a map of the 1898 railroad network, and maps of historical routes of exploration as sources of exogenous variation in the interstate highway network. Their main finding is that a 10% increase in kilometers of interstate highways causes an about 1.5% increase in employment over 20 years. Alternatively, a one standard deviation in initial roads causes a change in employment growth of about 15% over 20 years. This is a bit under two-thirds of the sample average growth rate.

Duranton and Turner (2012) also estimate a second equation in which they examine the effect on employment growth of changes in the stock of roads in the nearest large city. In the context of Figure 20.6, this corresponds to looking for an effect in the treated region from changes in the residual region. They find no effect. This regression, together with their main intracity regression, provides a tentative basis for concluding that roads cause employment growth in cities rather than simply rearranging employment across cities.

In a second exercise, Duranton et al. (2014) investigate the relationship between intercity trade flows in 2007 and the interstate highway network. Their unit of analysis is a US “commodity flow survey area”: a reporting unit somewhat larger than an MSA. They record the weight and value of pairwise trade flows between 69 such units and also aggregate flows in and out of each area by sector. On the basis of a method pioneered in Redding and Venables (2004) and Anderson and van Wincoop (2003), they develop two estimating equations. The first is a variant of the intercity regression and explains pairwise trade flows of weight and value as a function of pairwise interstate distance. The second is a variant of the intracity regression and predicts aggregate flows in and out of each city, by weight and value (irrespective of destination). In each case, they use the 1947 highway plan and the 1898 railroad network to derive instrumental variables. For the intracity regression, they also use instruments derived from routes of major explorations between 1530 and 1850. They arrive at three main findings. First, a 1% decrease in pairwise travel distance causes an about 1.4% increase in the value of pairwise trade and a 1.7% increase in its weight. Second, within-city highways affect the weight of exports, but not their value. Specifically, a 1% increase in the lane kilometers of within commodity flow survey area interstate highways causes an about 0.5% increase in the weight of exports but has no measurable effect on the value of exports. A 50-year panel of employment data confirms this result. Cities with more highways employ more people to make heavy manufactured goods, and conversely.

Finally, Duranton and Turner (2011) investigate the effect of the supply of roads and highways on the amount of driving in a city. More specifically, they conduct a version of the intracity regression. The outcome variable of interest is a measure of the total vehicle kilometers driven in a US MSA on particular road networks in a year and the explanatory variables of interest measure the extent of the road networks. They conduct this regression in levels, first differences and second differences. They also rely on maps of the 1947 highway plan, the 1898 railroad network, and routes of major expeditions of exploration

between 1530 and 1850 as sources of exogenous variation in MSA roads. They establish a “fundamental law of road congestion,” according to which driving increases by about 1% for each 1% increase in the stock of roadways, a finding that is robust across all of their specifications. They provide a rough decomposition of the sources of the marginal induced driving. About half comes from changes in individual behavior. Increases in commercial driving are less important. Migration in response to new roads and diversion of traffic from other networks appear to be least important. [Hsu and Zhang \(2012\)](#) replicate the analysis of [Duranton and Turner \(2011\)](#) using Japanese data. They arrive at the same conclusion: driving in Japanese cities increases by about 1% for each 1% increase in the extent of the road network.

While the above-mentioned articles are concerned with the relationship between overall traffic volumes and lane kilometers of roads, [Couture et al. \(2012\)](#) examine the determinants of driving speed in large US cities. Remarkably, their article is the first to estimate an econometric framework in which the supply and demand for travel are both explicitly modeled. The estimation results are used to construct a city-level index of driving speed and to undertake a welfare analysis of counterfactual changes in driving speed. Cities differ substantially in terms of driving speed, and the welfare gains from improvements in driving speed in the cities with the lowest driving speeds are found to be large. Taken together, these results are consistent with substantial deadweight losses from congestion.

Although most of the intracity literature is based on one of the three instrumental variables estimation strategies discussed above, the article by [Gonzalez-Navarro and Quintana-Domeque \(2013\)](#) is noteworthy for its use of a randomized experiment research design to examine the effects of road paving in Mexico. Homes in treatment streets that were paved experienced an increase in value of between 15% and 17% relative to those in control streets. The estimated rate of return to road pavement is 2% without taking into account externalities, but rises to 55% after incorporating externalities.

20.5.1.3 Subways and the internal organization of cities, and related other results

A large body of literature examines the effect of subways on the internal organization of cities. These articles typically consider a unit of analysis that is small relative to the city—for example, a census tract or zip code. The explanatory variable of interest is typically the distance to the subway. The outcome of interest is typically population or employment density, land prices, or ridership rates. That is, these articles perform a version of the inter-city regression (here inaptly named), Equation (20.32), at a subcity scale of analysis. As we discussed in [Sections 20.4.2 and 20.4.3](#), such regressions must overcome two problems: endogeneity and distinguishing between growth and reorganization.

The literature on subways is too large to survey exhaustively. We focus on three articles which provide, in our opinion, the best resolution to the endogeneity problem—[Gibbons and Machin \(2005\)](#), [Billings \(2011\)](#), and [Ahlfeldt et al. \(2014\)](#)—on two articles

showing that within-city roads are associated with qualitatively similar density gradients as subways—[Baum-Snow \(2007\)](#) and [Garcia-Lopez et al. \(2013\)](#)—and finally, on two articles which provide cross-city evidence of the effects of subways—[Baum-Snow and Kahn \(2005\)](#) and [Gordon and Willson \(1984\)](#). [Gibbons and Machin \(2005\)](#) and [Billings \(2011\)](#), in particular, provide more extensive surveys.

[Gibbons and Machin \(2005\)](#) conduct a difference-in-differences estimate of the inter-city estimation equation in order to evaluate the effect on London residential real estate prices of subway extension in the late 1990s. Their unit of observation is a “postcode unit,” an administrative unit containing 10–15 households. They observe real estate transactions by postcode unit before and after the Docklands light rail extension in south London. As a consequence of this extension, parts of their sample experience a decrease in the distance to a subway station. This makes a difference-in-differences estimate possible: they compare the change in real estate prices in postcodes that experienced changes in subway access with the change in real estate prices in postcodes that did not.

They find that for properties within 2 km of a station, a 1 km reduction in station distance causes an about 2% increase in real estate prices. Usefully, [Gibbons and Machin \(2005\)](#) compare their difference-in-differences estimate with a more conventional cross-sectional estimate. They find that estimates based on cross-sectional variation alone are $3\times$ as large as difference-in-difference estimates. This suggests that, as we might hope, subway station locations are not selected at random and more valuable land is more likely to receive subway service.

[Billings \(2011\)](#) and [Ahlfeldt et al. \(2014\)](#) also conduct difference-in-differences estimates of the effects of subways. For a newly opened light rail line in Charlotte, North Carolina, [Billings \(2011\)](#) finds that residential real estate prices within 1 mile of a station increase by about 4% for single-family homes and by about 11% for condominiums, and that light rail access has no effect on commercial property prices. [Ahlfeldt et al. \(2014\)](#) find that city blocks further than 250 m from a 1936 subway station experienced an about 13% smaller decrease in the price of floor space as a consequence of the division of Berlin than did those within 250 m. [Glaeser et al. \(2008\)](#) look at the effects of the New York city subway and find evidence that poor people move to be closer to subway stations.

Each of these three articles investigates the rate at which land rent declines with distance from a subway or light rail line. [Baum-Snow \(2007\)](#) and [Garcia-Lopez et al. \(2013\)](#) investigate how population density varies with distance to a highway. The unit of observation in [Baum-Snow \(2007\)](#) is a census tract. For each US census tract in a 1990 MSA, he observes the population density in 1970 and 1990 and the distance to an interstate highway. This allows him to estimate a variant of the intercity estimating equation for two cross sections and in first differences. He finds that a 10% decrease in the distance to a highway is associated with an about 0.13% increase in population density in 1970 and a slightly smaller increase in 1990. First-difference estimates are similar. [Garcia-Lopez et al. \(2013\)](#) arrive at similar estimates using Spanish data.

While each of these articles attempts to resolve the problem of endogenous placement of infrastructure, they do not provide a basis for determining whether subways cause growth or reorganization of nearby economic outcomes. In particular, they are unable to measure whether a change in a city's subway network affects city-level variables. In the context of [Figure 20.6](#), this would correspond to asking whether a change in treated unit infrastructure affects the level of an outcome in all three regions. This question, which is of obvious public policy interest, requires cross-city data describing subways and city-level outcomes—that is, data which allow the estimation of the intracity regression (Equation 20.31). Since subways are relatively rare, data of this sort are difficult to assemble, and we know of only two such efforts to date. The first, by [Gordon and Willson \(1984\)](#), constructs a single cross section of 52 cities that describes population density, subway passenger kilometers per year, and a handful of city-level control variables. In a simple cross-sectional estimate of ridership on density, they find a strong positive relationship. [Baum-Snow and Kahn \(2005\)](#) construct disaggregated panel data describing a panel of 16 US metropolitan areas with subways. In addition to describing the extent of each city's subway network, their data describe ridership commuting times. Overall, they find little evidence that US subway expansions elicit large increases in ridership.

20.5.2 Intercity infrastructure and the geographic organization of economic activity

We now turn our attention to the effect of infrastructure that connects a unit of observation, typically a county, to the rest of the world. This most often involves estimating a version of the intercity regression. We first describe results for high-income countries and then turn to results for low-income countries and historical data.

20.5.2.1 High-income countries

[Chandra and Thompson \(2000\)](#) consider the effect of the interstate highway system on a sample of 185 nonmetropolitan US counties that received a highway after 1969, and 391 neighboring nonmetropolitan counties that did not. By restricting their attention to nonmetropolitan counties, [Chandra and Thompson \(2000\)](#) hope to restrict their attention to counties that were treated with highways “accidentally,” and in particular, without regard for the effect of highways on the treated counties. This is the pioneering use of the inconsequential place approach to identification. Their outcome measures are aggregate annual earnings by county, year, and one-digit Standard Industrial Classification code, for all years from 1969 to 1993.

[Chandra and Thompson \(2000\)](#) estimate a distributed lag version of the intercity regression with county fixed effects. In particular, they include 24 dummies for the age of the highway connection in each year as explanatory variables. Their results are striking. They find a marginally positive 24-year effect of a highway connection on earnings in finance, insurance, real estate, transportation and public utilities, and retail

and services. They find that the effect on earnings in manufacturing and farming is marginally negative. Overall, the 24-year effect on earnings of a highway connection of a nonmetropolitan county is a 6–8% increase. The effect on untreated neighboring counties is approximately the opposite. Overall, untreated neighboring counties see a decrease in total earnings of between 1% and 3%. Note that [Chandra and Thompson \(2000\)](#) estimate two distinct equations. In the context of [Figure 20.6](#), the first predicts the effect of changes in infrastructure on the treated area, and the second predicts the effect of changes in infrastructure on neighboring untreated regions. Together, these two regressions are exactly what is required to distinguish between growth and reorganization. Importantly, [Chandra and Thompson \(2000\)](#) cannot reject the hypothesis that aggregate changes in earnings caused by a highway connection sum to zero across the whole sample of treated and neighboring counties.

[Michaels \(2008\)](#) considers a sample of 2000 counties in the United States that are more than 50% rural and had no highways in 1950—that is, the inconsequential place approach. He then identifies a subset of the interstate network constructed between 1959 and 1975 to serve intercity travel. His explanatory variable of interest is an indicator of whether a county is connected to this network at the end of the study period. He also relies on a planned route instrumental variable approach based on the 1947 highway plan. He considers a number of outcome variables, in particular, per capita earnings in trucking and retail sales, and the relative wages of skilled and unskilled workers. He finds that rural counties receiving highway connections experience about the same increase in trucking and retail earnings as [Chandra and Thompson \(2000\)](#) observe, the only two outcome variables common to the two articles. This is reassuring given the quite different identification strategies. He also finds that highways cause a small increase in the wage of skilled workers relative to that of unskilled workers.

In the first of two related, but methodologically quite different articles, [Redding and Sturm \(2008\)](#) consider the effect of the postwar partition of Germany on the organization of economic activity. They find that the population of German cities near the East–West border grew more slowly than that of those far from the border. That is, in response to an increase in the cost of travel between East Germany and West Germany, economic activity migrates away from the border region. [Duranton et al. \(2014\)](#) examine the effect of pairwise distance on pairwise trade of manufactured goods between US cities in 2007. They find that trade responds to highway distance rather than straight-line distance, that the effect of distance on trade is large, and that it is larger on the weight of goods than on their value. Unsurprisingly, [Duranton et al. \(2014\)](#) also find that trade by rail is less sensitive to distance than is trade by road. Curiously, [Duranton \(2014\)](#) replicates the analysis of [Duranton et al. \(2014\)](#) using data describing trade in Columbia rather than the United States. He reaches somewhat different conclusions: trade is less sensitive to distance, the value and the weight of trade are about equally sensitive to infrastructure, and the value of trade responds to infrastructure.

While most of the intercity literature has focused on roads, [Sheard \(2014\)](#) estimates the effects of airport infrastructure on relative sectoral employment at the metropolitan area level, using data from the United States. To address the potential endogeneity in the determination of airport sizes, the 1944 National Airport Plan is used as an instrument for the current distribution of airports. Airport size is found to have a positive effect on the employment share of tradeable services, controlling for overall local employment, but has no measurable effect on manufacturing or most nontradeable sectors. The effect of airport size on overall local employment is practically zero, suggesting that airports lead to specialization but not growth at the metropolitan area level. The implied elasticity of tradeable service employment with respect to airport size is approximately 0.22.

20.5.2.2 Low-income countries

[Donaldson \(2015\)](#) considers the effect of railroads on a sample of 235 “districts” covering the preponderance of India during the period from 1870 to 1930. He uses these data to estimate the intercity regression with district and year fixed effects. His outcome variable is the aggregate annual value of 17 agricultural crops per unit of district area. During this study period, agriculture accounted for about two-thirds of Indian GDP, and the 17 crops Donaldson considers accounted for 93% of the value of agricultural output. To investigate the probable endogeneity of railroads, Donaldson gathers data describing hypothetical planned railroad networks that were competitors to the realized network. He finds no difference in output between districts treated with planned networks and those not treated. This suggests that the realized network did not target the most productive districts.

Donaldson finds that districts with access to the railroad report about 17% higher real agricultural income per unit of district area than districts without railroads. Because Donaldson’s regression equation contains year and district effects, this means that a district treated with a railroad connection sees its income increase by 17% relative to untreated districts. This is a large effect. Over the course of the 1870–1930 study period, India’s real agricultural income increased by only about 22%, so a rail connection was equivalent to more than 40 years of economic growth.

In a related article, [Donaldson and Hornbeck \(2013\)](#) consider a sample of about 2200 counties in the continental United States between 1870 and 1890, a period of rapid rail expansion. They also perform a variant of the intercity regression, this time with county fixed effects, state-year fixed effects, and a cubic polynomial in latitude and longitude. The outcome variable of interest is the total value of a county’s agricultural land.

[Donaldson and Hornbeck \(2013\)](#) find that counties treated with rail access in a year experience a 34% increase in aggregate agricultural land rent relative to others in the same state and year. If the share of agricultural land in production stays approximately constant during their study period, then this implies the same effect on output, nominally larger than the corresponding estimate for India. With this said, the rate of growth in the United

States was much higher during this period, so a rail link was equivalent to only about 7.5 years of economic growth, as opposed to more than 40 years for Indian districts.

Beyond the inclusion of county fixed effects and other controls, [Donaldson and Hornbeck \(2013\)](#) do not have a strategy to deal with the endogeneity of rail access in the specification discussed above. Instead, they conduct an alternative regression where the explanatory variable of interest is a measure of market access. Their measure of market access results from a model similar to the one we described in [Section 20.3](#) and is well approximated by a “gravity” measure of population—that is, an inverse travel time weighted sum of county populations. They find that the effects of this measure are similar to those of the connection indicator. They also find that the effects of a restricted gravity measure, which excludes nearby counties, have a similar effect. That the two gravity measures have similar effects suggests that the effect of rail access on a county depends equally on rail access to places near and far away.

[Haines and Margo \(2008\)](#) conduct an analysis similar to that of [Donaldson and Hornbeck \(2013\)](#). They consider a sample of 655 counties in 12 US states and estimate the intercity regression in first differences. Their study period runs from 1850 to 1860, just before the 1870–1890 period that Donaldson and Hornbeck consider. They primarily consider the following outcome measures: share of urban population, agricultural wage, agricultural output per acre, and improved acreage share. Their measure of rail access is an indicator variable describing whether or not a rail line passes through a county in a year. They find that rail access is associated with a 10% increase in the share of a county’s improved acreage, a 3% increase in farm wages, no effect on output per improved acre, a small increase in service sector employment, and a 4% decrease in agricultural employment. In spite of the fact that [Haines and Margo \(2008\)](#) consider many of the same counties as [Donaldson and Hornbeck \(2013\)](#), and that the two study periods are adjacent, these results are much smaller than those obtained by [Donaldson and Hornbeck \(2013\)](#).

[Bogart \(2009\)](#) uses a sample of about 3000 English parishes and townships between 1692 and 1798 to estimate the intercity regression in first differences. His dependent variable is land rent per acre. His measure of transportation is an indicator of whether a parish or township is close to a turnpike, an improved road maintained by tolls. He also conducts an instrumental variables variant of the first-differences intercity regression, where he uses proximity to a major trade route as an instrument for the presence of a turnpike. This is a variant of the inconsequential places approach developed in [Chandra and Thompson \(2000\)](#). [Bogart \(2009\)](#) finds that a turnpike increases parish or township land rent by about 11% in first-difference estimates and by about 30% in instrumental variable estimates.

[Banerjee et al. \(2012\)](#) use county-level Chinese data to estimate the intercity regression with provincial and year fixed effects, and county-level controls. They consider a sample of 310 Chinese counties, for which they observe per capita GDP annually

from 1986 until 2006, a period when Chinese road and rail infrastructure expanded dramatically. They also consider a census of firms for a larger set of counties in a smaller number of years. To measure infrastructure, [Banerjee et al. \(2012\)](#) construct a hypothetical network connecting “treaty ports” to interior trading centers and use this network as an instrument. Again, this is a variant of the inconsequential places approach. Their measure of infrastructure is the distance from a county to a line in this hypothetical network, which predicts the proximity to both railroads and major highways.

Since [Banerjee et al. \(2012\)](#) have one instrument and two endogenous dependent variables, proximity to railroads and highways, they cannot separately identify the effects of roads and railroads. Instead, they present the results of an intercity regression in which the measure of transportation access is the distance to the hypothetical line. Therefore, as they acknowledge, these results are somewhat difficult to interpret. With this said, [Banerjee et al. \(2012\)](#) arrive at robust and interesting results. In particular, a 10% increase in the distance to a “line” causes an about 6% decrease in county GDP and has no effect on the growth of income. They find that the gradient for the density of firms is slightly steeper and that proximity to a line has no effect, or possibly a small negative effect, on the growth rate of firm density.

[Storeygard \(2012\)](#) uses a sample of 287 small cities in sub-Saharan Africa between 1992 and 2008 to estimate a first-differences variant of the intercity regression. This article is innovative in two regards. First, it uses “lights at night data” as a proxy measure for city GDP in small developing countries where data availability is limited.²¹ Second, to generate time series variance in transportation costs, he causes constant network distances to interact with a measure of the price of oil on international markets. As he observes, the validity of this approach hinges on the claim that, conditional on controls, oil prices do not affect city lights except through transportation costs. Thus, more specifically, for a sample of 287 small cities, [Storeygard \(2012\)](#) estimates a variant of the intercity regression where the outcome of interest is a measure of average annual light intensity for constant-boundary cities, and the measure of transportation costs is the interaction of network distance with annual average oil prices, city fixed effects, and variables to control for other possible channels through which oil prices might affect light intensity. [Storeygard \(2012\)](#) estimates that doubling the distance between a sample city and the primate port city causes an about 6% reduction in GDP, and that this is close to the effect of a quadrupling of fuel costs.

[Jedwab and Moradi \(2013\)](#) provide evidence regarding the intercity regression using rail construction in colonial sub-Saharan Africa, where over 90% of African railroad lines were built before independence. Colonial railroads are found to have strong effects on commercial agriculture and urban growth before independence. A number of

²¹ [Henderson et al. \(2012\)](#) pioneer the use of these data and show that they are highly correlated with national-level GDP, a result that [Storeygard \(2012\)](#) confirms at the subnational level.

identification strategies are used to provide evidence that these effects are causal, including placebo lines that were planned but not built and a version of the inconsequential units approach. Furthermore, using the fact that African railroads fell largely out of use after independence, owing to mismanagement and lack of maintenance, the article shows that colonial railroads had a persistent impact on cities. While colonial sunk investments (e.g., schools, hospitals, and roads) partly contributed to urban path dependence, the evidence suggests that railroad cities persisted because their early emergence served as a mechanism to coordinate contemporary investments for each subsequent period.

Faber (2015) also estimates a version of our intercity regression using a sample of about 1300 rural Chinese counties that are more than 50 km from a major city and that he observes in 1990, 1997, and 2006. For each county and year, he observes county-level GDP in three sectors—agriculture, industry, and services—as well as government expenditure. He also observes a rich set of county-level controls. His measure of infrastructure is the distance from the county centroid to the nearest segment of the trunk highway network, the limited access highway network that was substantially constructed during Faber’s study period. To resolve the probable endogeneity of the network placement, he relies on two hypothetical networks. The first resembles the hypothetical network developed by Banerjee et al. (2012). The second describes the cost-minimizing network to connect a set of major cities targeted by plans for the realized network. Faber (2015) finds that industrial GDP, total GDP, and government revenue all *decrease* with proximity to the network. This result, which appears robust, is without precedent in the literature. Every other implementation of this research design we survey arrives at the opposite conclusion—that is, that transportation infrastructure attracts (or creates) economic activity.

Ghani et al. (2013) use the inconsequential units approach to estimate the intercity regression for “The Golden Quadrilateral Project,” which upgraded the quality and width of 5846 km of roads in India. A difference-in-differences specification is used to compare non-nodal districts on the basis of their distance from the highway system. Positive treatment effects are found for non-nodal districts located 0–10 km from the Golden Quadrilateral that are not present in districts 10–50 km away, most notably for higher entry rates and increases in plant productivity.

20.6. DISCUSSION

20.6.1 Growth versus reorganization

Determining the extent to which the observed effects of infrastructure reflect changes in the level of economic activity versus a reorganization of existing activity is fundamental to understanding the effects of infrastructure and to policy analysis. The existing reduced-form literature generally does not provide a basis for separately identifying the two effects. In spite of this, we can suggest some tentative conclusions about the contributions of

growth and reorganization to the observed effects of infrastructure. These conclusions are based on comparisons between four sets of estimation results.

First, [Duranton et al. \(2014\)](#) examine the effect of within-city highways on the composition and value of intercity trade for US cities. They find that an increase in within-city highways causes cities to become more specialized in the production of heavy goods, but has at most small effects on the total value of trade. Here, the primary effect of within-city highways is to reorganize economic activity, not to create it.

Second, from the results in [Baum-Snow \(2007\)](#), [Garcia-Lopez et al. \(2013\)](#), and [Baum-Snow et al. \(2012\)](#), respectively, the effects of a one standard deviation increase in the number of radial highways cause the central city population to decrease by 14%, 5%, and 17% where secular rates of city population growth were 72%, 30%, and 55%. Thus, the transportation network causes reorganizations of cities that are large compared with the forces affecting them. On the other hand, [Duranton and Turner \(2012\)](#) find that a one standard deviation increase in within-city lane kilometers of interstate highways causes an about 15% increase in the population over 20 years. Happily, the samples of cities and years considered by [Baum-Snow \(2007\)](#) and [Duranton and Turner \(2012\)](#) substantially overlap. While the comparison is somewhat strained, it suggests that growth and reorganization are about equally important.

Third, [Banerjee et al. \(2012\)](#) conduct intercity regressions where the outcome variable is the level of GDP, and where it is the growth of GDP. They find that transportation infrastructure (really, their hypothetical network connecting treaty ports and interior trading centers) has important effects on the level of output, but not on its growth.

Fourth, and finally, [Chandra and Thompson \(2000\)](#) find that interstate highways increase firm earnings in US counties treated with interstate highways at the expense of their untreated neighbors. Summing over the treated and untreated counties, they cannot reject the hypothesis of no change.

While our evidence here is fragmentary, it suggests two conclusions. First, within large cities, relocation of economic activity in response to transportation infrastructure is at least as important as the creation of economic activity. This conclusion is broadly consistent with current estimates of agglomeration effects: if output increases by 2% with each doubling of city size, then even if infrastructure can double population size, we will see only small increases in productivity. Second, for nonurban counties, the primary effect of treatment with highways or railroads is to attract economic activity at the expense of more remote areas, with some variation by industry.

20.6.2 The effects of transportation infrastructure on economic activity

20.6.2.1 Invariance across economies

Quite different data underlie the three decentralization articles. [Baum-Snow \(2007\)](#) considers a 40-year study period and a US unit of observation with a mean population around

160,000. [Garcia-Lopez et al. \(2013\)](#) consider a 20-year study period and a Spanish unit of observation with a mean population around 120,000. [Baum-Snow et al. \(2012\)](#) consider a 20-year study period and a Chinese unit of observation with a population near 4 million. In spite of this, the three studies find remarkably similar effects of highways on the decentralization of the population from central cities to suburbs; 5% per ray for Spanish cities, 9% per ray for US cities, and 5% per ray for Chinese cities. That the effect of radial highways on population decentralization is so nearly the same in such different contexts suggests that the effects of infrastructure are not sensitive to the scale of the analysis or the details of the economies where the cities are located.²²

Other comparisons bolster this proposition. First, [Duranton and Turner \(2011\)](#) and [Hsu and Zhang \(2012\)](#) find, respectively, that a 1% increase in limited-access highways in a metropolitan area increases driving by 1% in US and Japanese metropolitan regions. Second, the effect of subways on land rent gradients appears to be about the same in suburban London as in Charlotte, North Carolina, while the effect of highway access on population density gradients appears similar in the United States and Spain.

Finally, with a few exceptions, there is broad agreement among the many articles that estimate the intercity regression: [Chandra and Thompson \(2000\)](#) find a 6–8% increase in firm earnings in counties adjacent to the interstate highway network; [Michaels \(2008\)](#) confirms the finding of [Chandra and Thompson \(2000\)](#) in the two industries where they overlap; [Donaldson \(2015\)](#) finds 17% higher real agricultural income for Indian districts with rail access; [Haines and Margo \(2008\)](#) find a 3% increase in farm wages for counties served by a railroad; [Bogart \(2009\)](#) finds an 11–30% increase in land rent for parishes served by a turnpike; [Banerjee et al. \(2012\)](#) find a 6% decrease in per capita income from doubling the distance to a hypothetical trade route; and [Storeygard \(2012\)](#) finds a 6% decrease in city light intensity from doubling the cost of travel to the primate city. [Donaldson and Hornbeck \(2013\)](#) and [Faber \(2015\)](#) are outliers, predicting a 34% increase in agricultural land rent for counties served by a railroad and a decrease in output for counties closer to a highway.

If we exclude the work of [Faber \(2015\)](#), and ignore the problem of comparing the gradient estimates of [Banerjee et al. \(2012\)](#) and [Storeygard \(2012\)](#) with discrete treatment effects in the others, these estimates are all within one order of magnitude.²³

²² It also suggests that the changes caused by radial highways may occur more rapidly than these 20- or 40-year study periods considered by extant research.

²³ [Banerjee et al. \(2012\)](#) consider the effect of the distance to a line, rather than an indicator for whether a line crosses a county. Therefore, their results cannot be compared directly with results based on treatment indicators. However, an average county in their sample is approximately 2000 km², the area of a square about 45 km on a side. Given this, doubling or quadrupling the distance from a county center to a line should usually be enough to remove an intersecting segment. This suggests that the effect of an indicator variable for line presence should be in the neighborhood of 6–12%.

Given the differences in the underlying economies that are the subject of these studies, this seems remarkable.

In sum, the literature suggests that transportation infrastructure has similar effects on the organization of economic activity across a range of countries and levels of development. More specifically, highways cause the decentralization of economic activity and an increase in its level in cities, highways cause a dramatic increase in driving, and highways and railroads cause an increase in economic activity in rural areas near highways. This conclusion is subject to four caveats. First, there is some disagreement among articles estimating the intercity regression. Second, although the methods and data used in these articles are similar, they are not identical, so comparisons between them need to be regarded with caution. Third, as we noted above, we do not have much basis for distinguishing growth from reorganization. Fourth, and finally, [Duranton et al. \(2014\)](#) and [Duranton \(2014\)](#) examine the effects of roads on trade in the United States and Columbia and find different effects.

20.6.2.2 Variability across activities and modes

While the literature surveyed above suggests a number of general results, it also provides suggestive evidence that different activities respond differently to changes in infrastructure. The three decentralization studies—[Baum-Snow \(2007\)](#), [Baum-Snow et al. \(2012\)](#), and [Garcia-Lopez et al. \(2013\)](#)—find that decreasing transportation costs leads the population to migrate to the lower-density periphery. Here, reductions in transportation costs reduce central city population density. [Baum-Snow et al. \(2012\)](#) find that manufacturing decentralizes along with the population.

Empirical results from the literature conducting intercity regressions also suggest heterogeneous responses by industries. [Chandra and Thompson \(2000\)](#) find different responses to the interstate highway access in rural counties by different sectors, a result confirmed in [Michaels \(2008\)](#). [Haines and Margo \(2008\)](#) find a shift of land into agriculture and of employment into services with rail access in nineteenth century United States counties. [Duranton et al. \(2014\)](#) find that US cities with more highways specialize in the production of heavier goods.

Finally, the gradient estimates in [Banerjee et al. \(2012\)](#) can be directly compared to within-city regressions estimating the effects of population density or land rent on proximity to a road—for example, [Baum-Snow \(2007\)](#) and [Garcia-Lopez et al. \(2013\)](#). This comparison suggests a much steeper gradient for economic activity near rural highways than near urban highways.

Broadly, these studies support the claim that the weight per unit value of output, land share of production, and sensitivity to agglomeration are all economically important determinants of how a firm or industry responds to changes in transportation infrastructure. The literature is as yet too incomplete to provide much insight into the relative

importance of these different factors. More speculatively still, highways may have larger effects on the organization of economic activity in rural areas than in cities.

20.6.2.3 Political economy of infrastructure allocation

As discussed above, a central issue in evaluating the effects of transportation improvements is that these improvements are not randomly assigned. Implicit evidence for the process through which transportation investments are assigned can be obtained by comparing the OLS coefficients for the intercity and intracity regressions (which capture the impact of transportation investments assigned through the existing political process) with the instrumental variable coefficients (which capture the impact of transportation investments assigned through quasi-experimental variation). In [Baum-Snow \(2007\)](#) and [Duranton and Turner \(2012\)](#), instrumental variable estimates are larger in magnitude than OLS estimates. This suggests that the equilibrium allocation process assigns roads to places growing more slowly than a randomly selected city. [Baum-Snow et al. \(2012\)](#) and [Garcia-Lopez et al. \(2013\)](#) find contrary results for China and Spain. Thus, conditional on the validity of their respective identification strategies, these articles point to implicit differences in the political economy of infrastructure funding across countries.

Further research is needed explicitly examining the political economy of transportation infrastructure investments. [Knight \(2002\)](#) examines the US Federal Aid Highway Program, over which the House Committee on Transportation and Infrastructure and the Senate Environment and Public Works Committee have jurisdiction. The article finds evidence that measures of the political power of state delegations affect the allocation of funds, including a state's proportion of members serving on the transportation authorization committee, the proportion of a state's representatives in the majority party, and the average tenure of a state's representatives. Federal highway grants are found to crowd out state highway spending, leading to little or no increase in net spending.

20.6.3 General equilibrium effects

Generally, studies of the effect of infrastructure on the internal organization of cities do not consider the role of market access. This occurs despite the fact that market access is a component of the theoretical precursor of both the intercity and intracity regression equations. This appears to rest on the assumption, usually implicit, that cities are small open units and that we can examine changes in their internal structure and level of economic activity without reference to other cities. In fact, [Duranton and Turner \(2012\)](#) make this small open city assumption explicit and attempt to test it by examining the effect on a target city of a change in the stock of roads in the nearest large city. While this is not a particularly satisfactory test, that they find no effect suggests that disregarding interactions between cities while studying the effect of transportation infrastructure on their internal workings is reasonable.

The problem of market access merits two further comments. First, for the purpose of examining pairwise trade flows, Redding and Venables (2004) develop a framework which allows the explicit estimation of market access and variants of estimating Equations (20.32) and (20.31) based on a two-step estimation procedure. It is this framework that Duranton et al. (2014) apply to their investigation of the effect of the interstate highway system on pairwise trade flows between US cities. Second, the extant empirical literature can be usefully divided into two classes. The first follows a long tradition of conducting city-level regressions that assume implicitly (or explicitly in the case of Duranton and Turner, 2012) that cities can be regarded as independent units. In this framework, what happens in each city is pinned down by the utility level in a residual rural sector. This implies that what happens in one city does not affect what happens in other cities. The second follows the trade or new economic geography literature—for example, Redding and Sturm (2008)—and supposes that the interactions between cities are important. An interesting area for further research is reconciling these two different approaches.

20.6.4 Structural estimation, general equilibrium, and welfare

The recent reduced-form literature has made important strides in identifying causal effects of infrastructure on economic activity in rural regions. Specifically, this literature estimates changes in economic activity by industry and changes in population for cities and rural regions. We are just beginning to investigate whether different modes of transportation have different effects. With this said, the existing literature provides at most suggestive evidence for the extent to which the observed effects of infrastructure reflect the reorganization or creation of economic activity. Progress on this issue appears to fundamentally require an econometric framework which is capable of dealing with general equilibrium effects such as the possibility that infrastructure moves activity from one unit to another.

In the remainder of this section, we discuss a number of studies that have used structural approaches to estimate intercity or intracity effects of transportation infrastructure. These studies highlight four main advantages of a structural approach. First, as discussed above, this approach enables general equilibrium effects to be captured. Second, a structural approach allows for the estimation or testing of specific economic mechanisms. Third, the estimated model can be used to quantify aggregate welfare effects (as, e.g., in Section 20.3.6). Fourth, the estimated model can be used to undertake counterfactuals and generate ex ante predictions for the effects of policies that have not yet been implemented (see, e.g., Section 20.3.8).

We begin with intercity studies. Redding and Sturm (2008) use the division of Germany after the Second World War and the reunification of Germany and in 1990 as a natural experiment to provide evidence in support of a quantitative model of

economic geography. As discussed above, in the aftermath of division, cities in West Germany close to the East German–West German border experienced a substantial decline in population growth relative to other West German cities, and the estimated treatment effect is larger for small cities than for large cities. In a multiregion extension of the [Helpman \(1998\)](#) model, the treatment effect of division on border cities depends on two parameter combinations that capture (a) the strength of agglomeration and dispersion forces and (b) the elasticity of trade with respect to distance. For plausible values of these parameter combinations, the model can account quantitatively for both the average treatment effect of division and the larger treatment effect for small cities than for large cities. Smaller cities are more adversely affected by division, because they are disproportionately dependent on markets in other cities.

[Donaldson \(2015\)](#) combines a general equilibrium trade model with archival data from colonial India to investigate the impact of India's vast railroad network. The empirical analysis is structured around an extension of the analysis of [Eaton and Kortum \(2002\)](#) to incorporate multiple agricultural commodities that shares some features with the theoretical framework developed in [Section 20.3](#). This model delivers four key theoretical predictions that are applied to the data. First, for goods that are traded between regions, price differences between those regions can be used to measure bilateral trade costs. Second, the model yields a gravity equation for bilateral trade flows that can be used to estimate the response of trade flows to trade costs. Third, railroads increase real income levels, as measured by the real value of land income per unit area. Fourth, as in the theoretical framework developed above, each location's trade share with itself is a sufficient statistic for welfare. Consistent with these predictions of the model, there is a strong and statistically significant estimated effect of railroads on real income levels, but this effect becomes statistically insignificant after controlling for the model's sufficient statistic of a region's own trade share. These results provide evidence that the estimated effects of railroads are capturing the goods trade mechanism emphasized in the model.²⁴

To quantify the intercity effects of road construction, [Duranton and Turner \(2012\)](#) develop a system of cities model that they use to derive a system of equations for employment and roadway growth that can be estimated empirically. Utility in each city depends on the quality of amenities, consumption of a numeraire composite good, distance traveled, and consumption of land. Productivity in producing the composite good increases with city employment through a standard agglomeration economy. The cost of travel per

²⁴ Transportation infrastructure may not only promote internal trade within countries (as considered here) but may also enable the interior regions of countries to participate in external (international) trade, as examined in [Fajgelbaum and Redding \(2013\)](#) using the natural experiment of Argentina's integration into the world economy in the late nineteenth century.

unit of distance decreases with the length of the roadway and increases with aggregate vehicle traffic through a standard congestion effect. Population mobility implies that utility in each city is equalized with utility in the outside alternative of a rural area. Equilibrium city size is determined by the willingness of residents to drive to the city center. Using equalization of utility between cities and rural areas, together with equilibrium in land and travel markets, one can express equilibrium city employment as a power function of the length of roadways. If a partial adjustment process is specified, according to which city employment growth is a function of the distance between a city's actual population and its equilibrium population, the model delivers the following equation for city employment growth:

$$n_{it+1} - n_{it} = A_1 + ar_{it} + \lambda n_{it} + c_1 x_i + \epsilon_{1it}, \quad (20.34)$$

where n_{it} is the logarithm of employment in city i at time t , r_{it} is the logarithm of the roadway, x_i are controls for city characteristics, and ϵ_{1it} is a stochastic error. If a similar partial adjustment process for road construction is specified, we obtain an analogous equation for the city roadway growth:

$$r_{it+1} - r_{it} = A_2 + \theta r_{it} + \eta n_{it} + c_2 x_i + \epsilon_{2it}, \quad (20.35)$$

where ϵ_{2it} is a stochastic error. The equilibrium logarithm of roadway length is assumed to depend on the logarithm of the city population, the city characteristics controls, x_i , and instruments, z_i , that satisfy the exclusion restriction of affecting the city population only through roadways:

$$r_{it} = A_3 + c_3 n_{it} + c_4 x_i + c_5 z_i + \epsilon_{3it}, \quad (20.36)$$

where ϵ_{3it} is a stochastic error. The identification assumptions for instrument validity are

$$c_5 \neq 0, \quad (20.37)$$

$$\text{Cov}(z, \epsilon_1) = 0, \quad (20.38)$$

$$\text{Cov}(z, \epsilon_2) = 0. \quad (20.39)$$

As discussed above, the instrumental variables estimates imply that a 10% increase in a city's stock of interstate highways causes an about 1.5% increase in its employment growth over 20 years. These instrumental variables estimates are somewhat larger than the OLS estimates. Therefore, an additional kilometer of highway allocated to a city at random is associated with a larger increase in employment or the population than for a road assigned to a city by the prevailing political process. These results are consistent with the view that the existing political process tends to assign highways to more slowly growing cities.

The intercity study of [Desmet and Rossi-Hansberg \(2013\)](#) highlights the way in which a general equilibrium model can be used to quantify the relative importance of different mechanisms and evaluate welfare effects. This paper develops a system of cities model that incorporates heterogeneity in productivity, amenities, and congestion costs as determinants of city sizes. Congestion costs are modeled as depending on city-specific transportation infrastructure. Data on US MSAs are used to estimate these city characteristics and decompose the variation in city sizes into their contributions. All three characteristics are important for explaining the observed city size distribution. Eliminating differences across cities in any one characteristic leads to large population reallocations but has small welfare effects (population reallocations of as much as 40% can have welfare gains of as little as 2%). This pattern of results is consistent with the idea that welfare is approximately equalized across cities in the initial equilibrium, in which case the envelope theorem implies small welfare effects from population reallocations. In contrast, when the same method is applied to Chinese cities, eliminating differences across cities in any one characteristic leads to both large population reallocations and large changes in welfare. These contrasting results between the two countries are consistent with urban policies in China playing an important role in determining relative city sizes and aggregate welfare.

The intercity study of [Allen and Arkolakis \(2013\)](#) also uses a structural approach to quantify alternative economic mechanisms and evaluate welfare effects. The article develops an Armington model of trade and factor mobility that incorporates both an economic component and a geographic component. The economic component combines the gravity structure of trade in goods with labor mobility to determine the equilibrium distribution of economic activity on a space with any continuous topography of exogenous productivity and amenity differences and any continuous bilateral trade costs. To incorporate the possibility of agglomeration and dispersion forces, the overall productivity and amenity in a location can endogenously depend on its population. The article provides general conditions for the existence, uniqueness, and stability of the spatial economic equilibrium. The geographic component of the model provides a microfoundation for bilateral trade costs as the accumulation of instantaneous trade costs along the least-cost route between locations. The model combining these economic and geographic components is used to estimate the topography of trade costs, productivities, and amenities in the United States. Geographic location is found to account for at least 20% of the spatial variation in US income. The construction of the US interstate highway system is estimated to increase welfare by 1.1–1.4%, which is substantially larger than its cost.

We now turn to intracity studies. Until recently, theoretical models of internal city structure were highly stylized, which limited their usefulness for empirical research. Much of the theoretical literature has focused on the monocentric city model, in which

firms are assumed to locate in a central business district and workers decide how close to live to this central business district.²⁵ Lucas and Rossi-Hansberg (2002) were the first to develop a model of a two-dimensional city, in which equilibrium patterns of economic activity can be nonmonocentric.²⁶ In their model, space is continuous and the city is assumed to be symmetric, so the distance from the center is a summary statistic for the organization of economic activity within the city. Empirically, however, cities are not perfectly symmetric because of variation in locational fundamentals, and most data on cities are reported for discrete spatial units such as blocks.²⁷

To address these challenges, Ahlfeldt et al. (2014) develop a quantitative theoretical model of internal city structure that allows for a large number of discrete locations within the city that can differ in their natural advantages for production, residential amenities, land supply, and transportation infrastructure. The model remains tractable and amenable to empirical analysis because of the stochastic formulation of workers' commuting decisions that follows Eaton and Kortum (2002) and McFadden (1974). The city is populated by an endogenous measure of $\bar{H}$ workers, who are perfectly mobile within the city and the larger economy. Workers experience idiosyncratic shocks to the utility they derive from each possible pair of residence and employment locations within the city. Workers choose their residence and employment locations and consumption of residential land and a tradeable final good to maximize their utility. This idiosyncratic formulation of utility yields a gravity equation for the probability of commuting from i to j (π_{ij}):

$$\pi_{ij} = \frac{T_{ij} \left(d_{ij} Q_i^{1-\beta} \right)^{-\epsilon} (B_i w_j)^\epsilon}{\sum_{r=1}^S \sum_{s=1}^S T_{rs} \left(d_{rs} Q_r^{1-\beta} \right)^{-\epsilon} (B_r w_s)^\epsilon}, \quad (20.40)$$

where T_{ij} is a Fréchet scale parameter that determines the average attractiveness of the bilateral commute from residence location i to employment location j , d_{ij} is the iceberg cost in terms of utility of commuting between i and j , Q_i is land prices. B_i denotes amenities at residential location i and w_j denotes wages at employment location j .

In this setting, transportation technology influences the organization of economic activity within the city through the matrix of bilateral commuting costs d_{ij} . Both residential amenities (B_i) and final goods productivity (A_j , which determines w_j) are characterized by agglomeration economies and hence depend on the transportation technology through the endogenous employment distribution. Ahlfeldt et al. (2014) use the division and reunification of Berlin as an exogenous shock to structurally estimate the strength of

²⁵ The classic urban agglomeration models of Alonso (1964), Mills (1967), and Muth (1969) impose a monocentric city structure. While Fujita and Ogawa (1982) and Fujita and Krugman (1995) allow for nonmonocentricity, they model one-dimensional cities on the real line.

²⁶ For an analysis of optimal urban land use policies in such a setting, see Rossi-Hansberg (2004).

²⁷ For empirical evidence regarding the extent to which the organization of economic activity within cities is indeed symmetric, see Brinkman (2013).

the model's agglomeration and dispersion forces and to show that the model can account quantitatively for the observed changes in city structure. The model also provides a framework that can be used to analyze the effects of other public policy interventions, such as transportation infrastructure investments that reduce commuting costs d_{ij} between pairs of locations.

Another structural intracity approach is that of [Combes et al. \(2012\)](#), which develops a method for estimating congestion costs (which depend on transportation technology) using land transactions data. The key insight behind this method is that residential mobility implies that urban (dis)amenities and commuting costs are ultimately reflected in land prices. A system of cities model is developed, in which each city is monocentric and workers face costs of commuting to the central business district. The model highlights that the elasticity of urban costs with respect to the city population is the product of three quantities: the elasticity of unit land prices at the city center with respect to the population, the share of land in housing, and the share of housing in consumption expenditure. With implementation of this method, the article's preferred estimates for these three elasticities are 0.72, 0.25, and 0.23, respectively. From the product of these three parameters, the preferred elasticity of urban costs with respect to the city population is 0.041, which is close to existing estimates of agglomeration economies in the form of the elasticity of city productivity with respect to the city population. This finding that cities operate near aggregate constant returns to scale suggests that the fundamental trade-off of spatial economics—between agglomeration economies and congestion costs—may play only a limited role in explaining the observed distribution of city sizes. This prediction is in turn consistent with the observation that cities of vastly different sizes exist and prosper.

20.7. CONCLUSION

To determine the causal effect of infrastructure on the spatial organization of economic activity, the central inference problem that researchers must overcome is that infrastructure is not assigned to locations at random, but is assigned rather on the basis of many of the same unobserved location characteristics that affect economic activity. The recent empirical literature is organized around three main approaches to this problem: planned route instrumental variable approach, historical route instrumental variable approach, and the inconsequential places approach. While these approaches remain open to criticism and refinement, they are about as good as can be hoped for in an environment where experiments seem implausible.

This literature suggests a number of tentative conclusions about the effects of infrastructure. Most studies estimate that population or employment density falls between 6% and 15% with a doubling of the distance to a highway or railroad (where railroads are the primary mode of transportation). Highways decentralize urban populations, and with less certainty, manufacturing activity. They may also lead to a complementary concentration

of services. Different sectors appear to respond differently to different modes of transportation and people respond differently from firms. The effects of infrastructure seem similar across countries at different stages of development.

While much effort has been directed to unraveling the problem of nonrandom assignment of infrastructure to places, much less has been directed to distinguishing between growth and reorganization. This distinction is clearly central to any understanding of the role of infrastructure and transportation costs in an economy. We suggest two approaches to resolving this problem. The first is a two-equation generalization of the current single-equation reduced-form models. The second relies on our structural model to resolve this problem. With this said, the literature does suggest that much of the estimated effect of transportation costs and infrastructure on the spatial organization of economic activity is probably due to reorganization rather than growth. Refining our understanding of this issue seems an obvious place for further research.

In addition to the largely reduced-form literature currently available, structural models of transportation costs and the spatial organization of economic activity are beginning to appear. Structural models have the important advantage of allowing for estimates of general equilibrium effects, such as the migration of economic activity in response to changes in transportation costs, on the basis of theoretically founded estimating equations. They also have obvious advantages for welfare and counterfactual analysis: available results suggest the importance of the “share of trade with self” as an indicator of welfare. With this said, there is disagreement in the literature on the fundamental assumptions underlying these models: in particular, whether we should think of cities as drawing people from the countryside or as competing with other cities for residents. Resolving this issue appears to be an important prerequisite for further progress.

Finally, the existing literature has devoted little attention, empirical or theoretical, to the dynamics of how transportation infrastructure affects economic development. In particular, there are few panel data studies conducting impulse response estimates. This seems to be an important, though difficult area for further research.

ACKNOWLEDGMENTS

We are grateful to Chang Sun and Tanner Regan for excellent research assistance. We also thank Nate Baum-Snow, Gilles Duranton, Will Strange, Vernon Henderson, and participants at the conference for the *Handbook of Regional and Urban Economics* for excellent comments and suggestions. Responsibility for any opinions, errors and omissions lies with the authors alone.

REFERENCES

- Ahlfeldt, G.M., Redding, S.J., Sturm, D.M., Wolf, N., 2014. The economics of density: evidence from the Berlin Wall. Technical report. NBER Working paper 20354.
- Allen, T., Arkolakis, C., 2013. Trade and the Topography of the Spatial Economy. Yale University. Processed.

- Alonso, W., 1964. *Location and Land Use: Toward a General Theory of Land Rent*. Harvard University Press, Cambridge, MA.
- Anderson, J.E., van Wincoop, E., 2003. Gravity with gravitas: a solution to the border puzzle. *Am. Econ. Rev.* 93, 170–192.
- Ángel García-López, M., 2012. Urban spatial structure, suburbanization and transportation in Barcelona. *J. Urban Econ.* 72, 176–190.
- Atalay, E., Hortacsu, A., Syverson, C., 2013. *Vertical Integration and Input Flows*. University of Chicago. Processed.
- Banerjee, A., Duflo, E., Qian, N., 2012. On the road: transportation infrastructure and economic growth in China. Technical report. NBER Working paper No. 17897.
- Baum-Snow, N., 2007. Did highways cause suburbanization? *Q. J. Econ.* 122, 775–805.
- Baum-Snow, N., Kahn, M.E., 2005. Effects of urban rail transit expansions: evidence from sixteen cities, 1970–2000. *Brook. Wharton Pap. Urban Affa.* 2005 1 (4), 147–197.
- Baum-Snow, N., Brandt, L., Henderson, J.V., Turner, M.A., Zhang, Q., 2012. *Roads, Railroads and Decentralization of Chinese Cities*. University of Toronto. Processed.
- Bernhofen, D.M., El-Sahli, Z., Kneller, R., 2013. Estimating the Effects of Containerization on World Trade. University of Nottingham. Processed.
- Billings, S.B., 2011. Estimating the value of a new transit option. *Reg. Sci. Urban Econ.* 41, 525–536.
- Bleakley, H., Lin, J., 2012. Portage: path dependence and increasing returns in U.S. history. *Q. J. Econ.* 127, 587–644.
- Bogart, D., 2009. Turnpike trusts and property income: new evidence on the effects of transport improvements and legislation in eighteenth-century England. *Econ. Hist. Rev.* 62, 128–152.
- Brinkman, J.C., 2013. *Transportation Technologies, Agglomeration, and the Structure of Cities*. Philadelphia Federal Reserve Bank. Processed.
- Bureau of Labour Statistics, 2012. Consumer expenditure survey: average annual expenditures and characteristics of all consumer units. www.bls.gov (accessed 2013-06-01).
- Bureau of Transportation Statistics, 2012a. National transportation statistics: table 1–49, US ton-miles of freight. www.rita.dot.gov (accessed 2013-05-21).
- Bureau of Transportation Statistics, 2012b. National transportation statistics: table 3–23, employment in for-hire transportation and selected related industries. www.rita.dot.gov (accessed: 2013-05-02).
- Census Bureau, 2009. Commuting (journey to work): travel time tables. www.census.gov (accessed 2013-06-01).
- Chandra, A., Thompson, E., 2000. Does public infrastructure affect economic activity? Evidence from the rural interstate highway system. *Reg. Sci. Urban Econ.* 30, 457–490.
- China Data Online, 2010. China statistical yearbook 2010: table 16–9, freight ton-kilometers. www.chinadataonline.org (accessed 2013-05-23).
- China Data Online, 2011. China statistical yearbook 2011: tables 10–5 (urban) and 10–25 (rural). www.chinadataonline.org (accessed 2013-06-18).
- Clark, X., Dollar, D., Micco, A., 2004. Port efficiency, maritime transport costs, and bilateral trade. *J. Dev. Econ.* 75, 417–450.
- Combes, P.P., Duranton, G., Gobillon, L., 2012. *The Costs of Agglomeration: Land Prices in French Cities*. University of Pennsylvania. Processed.
- Cosar, K., Fajgelbaum, P., 2013. *Internal Geography, International Trade, and Regional Specialization*. University of California, Los Angeles. Processed.
- Couture, V., Duranton, G., Turner, M.A., 2012. *Speed*. University of Toronto. Processed.
- Criscuolo, C., Martin, R., Overman, H.G., Reenen, J.V., 2012. *The causal effects of an industrial policy*. Technical report. CEP Discussion Paper No 1113.
- Cronon, W., 1991. *Nature's Metropolis: Chicago and the Great West*. W. W Norton & Co, New York, NY.
- Davis, M.A., Ortalo-Magne, F., 2011. Household expenditures, wages, rents. *Rev. Econ. Dyn.* 14, 248–261.
- Davis, D.R., Weinstein, D.E., 2002. Bones, bombs, and break points: the geography of economic activity. *Am. Econ. Rev.* 92, 1269–1289.

- Dekle, R., Eaton, J., Kortum, S., 2007. Unbalanced trade. *Am. Econ. Rev.* 97, 351–355.
- Department for Transport, 2012. Table tsqb0306: greenhouse gas emissions by transport mode. www.gov.uk (accessed 2013–06–08).
- Desmet, K., Rossi-Hansberg, E., 2013. Urban accounting and welfare. *Am. Econ. Rev.* 103 (6), 2296–2327.
- Donaldson, D., 2015. Railroads of the Raj: estimating the impact of transportation infrastructure. *Am. Econ. Rev.* (forthcoming).
- Donaldson, D., Hornbeck, R., 2013. Railroads and american economic growth: a “market access” approach. Technical report. NBER Working paper No. 19213.
- Duranton, G., 2014. Roads and Trade in Columbia. University of Pennsylvania. Processed.
- Duranton, G., Morrow, P., Turner, M.A., 2014. Roads and trade: evidence from the US. *Rev. Econ. Stud.* 81, 681–724.
- Duranton, G., Turner, M.A., 2011. The fundamental law of road congestion: evidence from US cities. *Am. Econ. Rev.* 101, 2616–2652.
- Duranton, G., Turner, M.A., 2012. Urban growth and transportation. *Rev. Econ. Stud.* 79, 1407–1440.
- Eaton, J., Kortum, S., 2002. Technology, geography, and trade. *Econometrica* 70, 1741–1780.
- Eurofound, 2000. Third European survey on working conditions 2000: figure 37. www.eurofound.europa.eu (accessed 2013–06–17).
- Eurostat, 2009. Mean consumption expenditure by detailed COICOP level. www.epp.eurostat.ec.europa.eu (accessed 2014–05–18).
- Eurostat, 2010. Transport database. www.epp.eurostat.ec.europa.eu (accessed 2013–05–01).
- Eurostat, 2012. Extra EU27 trade since 1999 by mode of transport (NSTR). www.epp.eurostat.ec.europa.eu (accessed 2013–06–17).
- Faber, B., 2015. Trade integration, market size, and industrialization: evidence from China’s National Trunk Highway system. *Rev. Econ. Stud.* (forthcoming).
- Fajgelbaum, P., Redding, S., 2013. External Integration, Structural Transformation and Economic Development: Evidence from Argentina. Princeton University. Processed.
- Fogel, R., 1964. Railroads and American Economic Growth: Essays in Econometric History. Johns Hopkins University Press, Baltimore, MD.
- Fujita, M., Krugman, P., 1995. When is the economy monocentric? Von Thünen and Chamberlain unified. *Reg. Sci. Urban Econ.* 25, 505–528.
- Fujita, M., Ogawa, H., 1982. Multiple equilibria and structural transition of non-monocentric urban configurations. *Reg. Sci. Urban Econ.* 12, 161–196.
- Fujita, M., Krugman, P., Venables, A., 1999. The Spatial Economy: Cities, Regions, and International Trade. MIT Press, Cambridge, MA.
- Garcia-Lopez, M.A., Holl, A., Viladecans-Marsal, E., 2013. Suburbanization and Highways: When the Romans, the Bourbons and the First Cars Still Shape Spanish Cities. Universitat Autònoma de Barcelona and IEB.
- Ghani, E., Goswami, A.G., Kerr, W.R., 2013. Highway to success in India: The impact of the Golden Quadrilateral project for the location and performance of manufacturing. Technical report. World Bank. Policy Research Working paper 6320.
- Gibbons, S., Machin, S., 2005. Valuing rail access using transport innovations. *J. Urban Econ.* 57, 148–169.
- Glaeser, E.L., Kohlhase, J.E., 2004. Cities, regions and the decline of transport costs. *Pap. Reg. Sci.* 83, 197–228.
- Glaeser, E.L., Ponzetto, G.A.M., 2013. Did the death of distance hurt Detroit and help New York? Technical report. NBER Working paper 13710.
- Glaeser, E.L., Kahn, M.E., Rappaport, J., 2008. Why do the poor live in cities? The role of public transportation. *J. Urban Econ.* 63, 1–24.
- Gonzalez-Navarro, M., Quintana-Domeque, C., 2013. Roads to Development: Experimental Evidence from Urban Road Pavement. University of Toronto. Processed.
- Gordon, P., Willson, R., 1984. The determinants of light-rail transit demand—an international cross-sectional comparison. *Transp. Res. A Gen.* 18, 135–140.
- Haines, M.R., Margo, R.A., 2008. Railroads and local economic development: the United States in the 1850s. In: Rosenbloom, J.L. (Ed.), *Quantitative Economic History: The Good of Counting*. Routledge, London, UK, pp. 78–99.

- Head, K., Mayer, T., 2013. What separates us? Sources of resistance to globalization. *Can. J. Econ.* 46, 1196–1231.
- Helpman, E., 1998. The size of regions. In: Pines, D., Sadka, E., Zilcha, I. (Eds.), *Topics in Public Economics: Theoretical and Applied Analysis*. Cambridge University Press, Cambridge, pp. 33–54.
- Henderson, J.V., Storeygard, A., Weil, D.N., 2012. Measuring economic growth from outer space. *Am. Econ. Rev.* 102, 994–1028.
- Hillberry, R., Hummels, D., 2008. Trade responses to geographic frictions: a decomposition using micro-data. *Eur. Econ. Rev.* 52, 527–550.
- Hsu, W.T., Zhang, H., 2012. The fundamental law of highway congestion: evidence from national expressways in Japan. Technical report. Working paper, Department of Economics, The National University of Singapore.
- Hummels, D., 1999. Towards a geography of trade costs. Technical report. GTAP Working paper No. 17.
- Jedwab, R., Moradi, A., 2013. Transportation Technology and Economic Change: Evidence from Colonial Railroads and City Growth in Africa. George Washington University. Processed.
- Knight, B., 2002. Endogenous federal grants and crowd-out of state government spending: theory and evidence from the federal highway aid program. *Am. Econ. Rev.* 92, 71–92.
- Krugman, P.R., 1991. Increasing returns and economic geography. *J. Pol. Econ.* 99 (3), 483–499.
- Limao, N., Venables, A.J., 2001. Infrastructure, geographical disadvantage, transport costs and trade. *World Bank Econ. Rev.* 15, 451–479.
- Lucas Jr., R.E., Rossi-Hansberg, E., 2002. On the internal structure of cities. *Econometrica* 70, 1445–1476.
- Martincus, C.V., Carballo, J., Cusolito, A., 2012. Routes, Exports, and Employment in Developing Countries: Following the Trace of the Inca Roads. World Bank. Processed.
- Mayer, T., Mayneris, F., Py, L., 2013. The impact of urban enterprise zones on establishments' location decisions: evidence from French ZFUs. Technical report. CEPR Discussion Paper 9074.
- Maystadt, J.F., Duranton, G., 2014. The Development Push of Refugees: Evidence from Tanzania, Wharton. Processed.
- McFadden, D., 1974. The measurement of urban travel demand. *J. Publ. Econ.* 3, 303–328.
- Michaels, G., 2008. The effect of trade on the demand for skill—evidence from the Interstate Highway System. *Rev. Econ. Stat.* 90, 683–701.
- Michaels, G., Rauch, F., Redding, S.J., 2012. Urbanization and structural transformation. *Q. J. Econ.* 127, 535–586.
- Mills, E.S., 1967. An aggregative model of resource allocation in a metropolitan area. *Am. Econ. Rev. (Pap. Proc.)* 57, 197–210.
- Moses, L.N., 1958. Location and the theory of production. *Q. J. Econ.* 72, 259–272.
- Moses, L.N., Williamson, H.F., 1963. Value of time, choice of mode, and the subsidy issue in urban transportation. *J. Pol. Econ.* 71, 247–264.
- Muth, R.F., 1969. *Cities and Housing*. University of Chicago Press, Chicago, IL.
- Natural Resources Canada, 2012. Comprehensive energy use database table 8: GHG emissions by transportation mode. www.oee.nrcan.gc.ca (accessed 2013-06-03).
- Neumark, D., Simpson, H., 2014. Place-based policies. In: Duranton, G., Strange, W., Henderson, V. (Eds.), *Handbook of Regional and Urban Economics*, vol. 5. Elsevier, Amsterdam, pp. 1197–1288.
- North American Transportation Statistics, 2012a. Table 2–3: employment in transportation and related industries. www.nats.sct.gob.mx (accessed 2013-05-23).
- North American Transportation Statistics, 2012b. Table 4–2: greenhouse gas emissions by transportation sector. www.nats.sct.gob.mx (accessed 2013-06-03).
- North American Transportation Statistics, 2012c. Table 5–2: domestic freight activity. www.nats.sct.gob.mx (accessed 2013-05-23).
- North American Transportation Statistics, 2012d. Table 7–1: international merchandise trade by mode. www.nats.sct.gob.mx (accessed 2013-06-03).
- OECD Stat Extracts, 2012. National accounts: table 5, final consumption expenditure of households. www.stats.oecd.org (accessed 2013-06-08).
- Office of National Statistics, 2009. Labour force survey: 11.6 time taken to travel to work by workplace. www.ons.gov.uk (accessed 2013-06-08).

- Parry, I.W.H., Walls, M., Harrington, W., 2007. Automobile externalities and policies. *J. Econ. Liter.* 45, 373–399.
- Ramondo, N., Rodriguez-Clare, A., Saborio, M., 2012. Scale Effects and Productivity: Does Country Size Matter? University of California at Berkeley. Processed.
- Redding, S., 2012. Goods trade, factor mobility and welfare. NBER Working paper.
- Redding, S., Sturm, D., 2008. The costs of remoteness: evidence from German division and reunification. *Am. Econ. Rev.* 98, 1766–1797.
- Redding, S., Venables, A.J., 2004. Economic geography and international inequality. *J. Int. Econ.* 62, 63–82.
- Redding, S.J., Sturm, D.M., Wolf, N., 2011. History and industry location: evidence from German airports. *Rev. Econ. Stat.* 93, 814–831.
- Rossi-Hansberg, E., 2004. Optimal urban land use and zoning. *Rev. Econ. Dyn.* 7, 69–106.
- Sarvimäki, M., Uusitalo, R., Jäntti, M., 2010. The long-term effects of forced migration. Aalto University. Processed.
- Schafer, A., 2000. Regularities in travel demand: an international perspective. *J. Transp. Stat.* 3, 1–31.
- Sheard, N., 2014. Airports and urban sectoral employment. *J. Urban Econ.* 80, 133–152.
- Small, K.A., Verhoef, E.T., 2007. *The economics of urban transportation*. Routledge, New York, NY.
- Statistics Canada, 2010. Commuting to work: results of the 2010 general social survey, table 1. www.statcan.gc.ca (accessed 2013-06-03).
- Statistics Canada, 2012. Cansim: table 384-0041, detailed household final consumption expenditure. www.statcan.gc.ca (accessed 2013-06-03).
- Storeygard, A., 2012. Farther on down the road: transport costs, trade and urban growth in Sub-Saharan Africa. Tufts University. Working paper.
- Turcotte, M., 2005. The time it takes to get to work and back: table 1. www.publications.gc.ca (accessed 2013-06-03).

CHAPTER 21

Cities in Developing Countries: Fueled by Rural–Urban Migration, Lacking in Tenure Security, and Short of Affordable Housing

Jan K. Brueckner*, Somik V. Lall†

*Department of Economics, University of California, Irvine, CA, USA

†Urban Development and Resilience Unit, Sustainable Development Network, World Bank, USA

Contents

21.1. Introduction	1400
21.2. The Empirical Aspects of Rural–Urban Migration	1402
21.2.1 Early studies	1404
21.2.2 More recent research	1404
21.2.2.1 <i>Adding covariates to combat selection bias</i>	1405
21.2.2.2 <i>Introducing access to public services in migration decisions</i>	1405
21.2.2.3 <i>Climate migrants</i>	1407
21.2.2.4 <i>Income risk as a factor governing migration</i>	1408
21.3. Models of Migration and City Sizes in Developing Countries	1409
21.3.1 The Harris–Todaro model	1410
21.3.2 Urban cost of living as an equilibration mechanism	1412
21.3.3 Hybrid models	1414
21.3.4 Dynamics	1415
21.3.5 Socially optimal city sizes	1416
21.3.6 Empirical evidence	1420
21.4. Tenure Insecurity: A Hallmark of Housing Markets in Developing Countries	1422
21.4.1 A continuum of property rights: Mali and Vietnam	1423
21.4.2 The economics of squatting: Theory	1425
21.4.2.1 <i>The Jimenez model</i>	1426
21.4.2.2 <i>The Hoy–Jimenez and Turnbull Models</i>	1428
21.4.2.3 <i>The Brueckner–Selod model</i>	1430
21.4.2.4 <i>The Shah model</i>	1432
21.4.3 The economics of squatting: Empirical work	1434
21.4.4 Avoiding the formation of squatter settlements or relocating them	1438
21.5. Provision of Affordable Housing in Developing Countries	1439
21.5.1 Investment in housing	1439
21.5.1.1 <i>The responsiveness of supply</i>	1439
21.5.1.2 <i>Some determinants of housing affordability</i>	1441

21.5.2	Land-management policies	1443
21.5.2.1	<i>Land-consumption standards</i>	1443
21.5.2.2	<i>Density regulations</i>	1444
21.5.2.3	<i>Urban growth boundaries</i>	1446
21.5.3	Investment in complementary infrastructure and services	1447
21.6.	Conclusion	1448
	Appendix	1450
	Acknowledgments	1451
	References	1451

Abstract

This chapter surveys and synthesizes existing research on urbanization and housing in developing countries. The goal is to provide a unified overview of the principal urban issues that arise in developing countries, painting a coherent picture that can provide a starting point for policy analysis. The chapter covers empirical work on rural–urban migration, theoretical research on migration and city-size determination, theoretical and empirical work on tenure security and squatting, and the issue of housing affordability.

Keywords

Rural–urban migration, Squatting, Property rights, Housing affordability, Land-use regulations

JEL Classification Codes

R1, O18

21.1. INTRODUCTION

According to the [World Bank \(2013\)](#), developing countries must prepare to house an additional 2.7 billion people between now and 2050, as migrants move in unprecedented numbers from rural areas to pursue their hopes and aspirations in cities. While many migrants seek the jobs that cities can offer, others come in search of public services not available in rural areas or seeking refuge from climate shocks that increase the vulnerability of rural livelihoods. Surging populations will place intense pressure on basic services and urban infrastructure at a time when developing-country cities still lack the resources and institutions to provide all the new arrivals with access to jobs, housing, and basic services. In fact, international evidence highlights that much urbanization takes place early in development, before countries have reached middle incomes ([World Bank, 2008](#)). As a result, migrants often settle in slums and squatter areas, which may provide inadequate shelter and lack security of tenure. The United Nations estimates that at least 860 million people live in slums across the developing world, with the number of slum dwellers growing by 6 million each year from 2000 to 2010 ([United Nations Habitat, 2012](#)). In sub-Saharan Africa, slum populations are estimated to be growing at 4.5% per annum, projected to double every 15 years ([Marx et al., 2013](#)).

While such settlements are not new in the history of rapidly growing cities, their persistence is as much an outcome of policy failures that restrict the supply of affordable housing as it is a result of surging urban populations, as argued by [Hammam \(2013\)](#). Improving the living conditions of slum dwellers by improving access to affordable housing is thus an imperative for urban policy. Housing also plays an important role in economic development as it is predominantly a private, household investment and usually the largest asset for most households. Some estimates suggest that housing often accounts for half a country's tangible capital stock and one-fifth to one-third of gross fixed capital formation and generates services equal to 10–30% of consumption expenditure ([Malpezzi, 2012](#)).

The purpose of this chapter is to survey and synthesize existing research on these aspects of urbanization in developing countries. The goal is to provide a unified overview of the principal urban issues that arise in developing countries, painting a coherent picture that can provide a starting point for policy analysis. We begin in [Section 21.2](#) with a treatment of the main driving force behind urbanization in developing countries: rural–urban migration. Traditional empirical work on such migration is summarized, and the discussion highlights new empirical research designed to improve on past studies. This new work studies the effect of public service availability on migration decisions, the role of income risks at the migration destination and of environmental risks (such as rainfall variability) in the home region, and the effect of using a broader set of individual migrant and household characteristics in measuring the migration impacts of traditional economic variables.

With this background in the empirical aspects of migration, [Section 21.3](#) turns to a discussion of theoretical models that capture the role of rural–urban migration in determining city sizes in developing countries. The famous Harris–Todaro model is the starting point, but the discussion then considers models where urban cost-of-living escalation, in the form of higher housing prices, chokes off rural–urban migration, providing an equilibrating mechanism different from the familiar unemployment channel in the Harris–Todaro model. Hybrid models, where both unemployment and cost-of-living forces jointly serve to equilibrate city sizes, are also explained. Our attention then shifts to the social optimality of the city sizes generated by rural–urban migration, identifying forces that may lead to inefficient migration equilibria, where cities are either too large or too small.

[Section 21.4](#) focuses on insecure housing tenure, a major feature of housing markets in developing countries. Insecure tenure derives from “land-rights insecurity,” or the insecurity of property rights to land and housing. The initial discussion emphasizes that housing markets in developing countries frequently offer a “continuum” of property rights, ranging from a complete lack of legal tenure security in squatter settlements, to moderate security levels among households with some legal documentation but no formal title, to the full security usually associated with a title. The bulk of this section,

however, focuses on one end of the tenure-security continuum by providing an extensive survey of research on urban squatting in developing countries. The chapter's emphasis on squatting is justified by the existence of a sizable theoretical and empirical literature on the topic, which represents one of the most substantial and coherent bodies of work by economists on urban phenomena in the developing world. Theoretical work on squatting is considered first, with the discussion reviewing the structure of several alternative models of the squatting phenomenon. Attention then shifts to empirical work on squatting, which attempts to measure the dollar cost of tenure insecurity as well as its various effects (including low housing investment by squatters).

Lack of affordable housing is another major problem in developing countries, and this issue is the subject of [Section 21.5](#). The discussion starts with an appraisal of the housing-supply response to urbanization that is driven by rural–urban migration, recognizing that responsive supply is crucial in maintaining affordability in the face of population pressure. Our attention then turns to other factors affecting affordability, including subsistence pressure that prevents households from diverting an appreciable share of their budgets away from food expenditure. The role of land-management policies in limiting the supply of affordable housing is also discussed. The policies include minimum-lot-size regulations, which are prevalent in many developing countries, other restrictions such as building-height limits that reduce densities and thus cut housing supply, and urban growth boundaries (UGBs), which also depress the supply of land and raise prices, with unfavorable effects on affordability. The discussion concludes by considering the role of urban public services in making urban living affordable to migrant households. The chapter's overall conclusions are presented in [Section 21.6](#).

21.2. THE EMPIRICAL ASPECTS OF RURAL–URBAN MIGRATION

Economic growth in less-developed countries has been accompanied by explosive urbanization. United Nations data show that the annual growth rate of the urbanized population in developing countries ranged between 2.7% and 4.2% across the years between 1950 and 2010, far outpacing the urban growth rates in developed countries, which ranged between 0.6% and 2.4%. Because of its faster growth, the urbanized share of the population in developing countries rose rapidly from 17.6% in 1950 to 46% in 2010, but still falls well short of the 77.5% share in developed countries for 2010. The explosive urban growth in developing countries has also created very large cities. In 2014, only four of the world's largest 15 cities are in developed countries.

Although high birth rates make the natural increase of the population an important source of city growth in developing countries, rural–urban migration is an even more significant force driving the urbanization process. This migration has been the focus of substantial research by economists, demographers, and other social scientists. The research, which is epitomized in the early studies of [Fields \(1982\)](#) and [Schultz](#)

(1982), shows that rural–urban migration responds to economic incentives in a natural fashion.¹

Migration depends on forces that “pull” migrants to their destinations as well as forces that “push” them to leave their origins. Better economic opportunities in cities, partly due to agglomeration economies, are a major pull factor, often providing the main motivation for internal migration. But people are also pushed off their land by severe declines in agriculture, by the pressures of population growth, and by environmental changes that make cultivation no longer viable. Historically, droughts have had sudden and prolonged impacts on the population distribution in developing countries, particularly in sub-Saharan Africa and South Asia.² Conflict has also pushed people to migrate in sub-Saharan Africa and in many other developing regions.

In many low- and middle-income countries, another important push factor propels internal migration: the lack of adequate basic services in rural areas or in economically lagging regions. To a large extent, this topic has been overlooked in empirical analysis of migration decisions. In reality, however, the location of schools, health care centers, hospitals, and other public amenities can significantly influence migration. In Africa, disparities in school enrollment and neonatal care across cities, towns, and villages are attributable to the near absence of schools and health facilities in outlying areas, and they constitute a force driving migrants toward cities. Evidence from Central Asia shows that, in the isolated parts of Tajikistan, schools are inadequately heated, drinking water is scarce, and garbage and sewage removal is lacking, again making migration attractive.³

Even though many forces provide strong incentives for migration, cultural and linguistic factors tend to make labor less mobile than capital, which moves quickly in a globalizing economy. Moreover, relative to capital, labor is subject to more political restrictions and to explicit and implicit barriers.⁴ For example, the household registration (or hukuo) system in China constitutes a major barrier to rural–urban migration since migrants without hukuo rights do not qualify for public education or health benefits. Recent research for China indicates that removing such mobility restrictions would reallocate labor across areas, reduce wage differences, and lower income inequality (Whalley and Zhang, 2004).

¹ See Lall et al. (2006a) for a comprehensive survey of the literature on rural–urban migration in developing countries.

² See Iliffe (1995) for the historical impact of drought on the population distribution in Africa, Bryceson (1999) for the Sahel and Sudan, and Hardoy and Satterthwaite (1989) for Mauritania. Wandschneider and Mishra (2003), cited in Deshingkar and Grimm (2004), provide evidence regarding the drought-induced migration of 60,000 people out of Bolangir, in the Indian state of Orissa, in 2001.

³ See Sahn and Stifel (2003) and Anderson and Pomfret (2005).

⁴ See World Bank (2009).

21.2.1 Early studies

Harris and Todaro (1970) produced the seminal conceptual work on rural–urban migration in developing countries. Their framework diverges from a standard model with full employment and flexible wages by introducing a politically determined minimum urban wage, which exceeds the agricultural wage and generates urban unemployment. Migration occurs as a response to the rural–urban difference in expected earnings, which depends on the city’s unemployment rate, and migration proceeds until expected earnings are equalized. This dual-economy model demonstrates the rationality of migration from rural to urban areas despite the possibility of urban unemployment.

The model generates a potential paradox—namely, that one extra minimum-wage job could induce more than one agricultural worker to migrate to the urban area, hence increasing the unemployment rate. Todaro (1976) asks whether the conditions leading to this paradox hold empirically. To this end, he analyzes developing-country data and estimates rural–urban migration elasticities, finding that their values are in a range where the paradoxical result seems to hold. On the basis of this evidence, Todaro (1976) concludes that “there would thus appear to be no strictly urban solution to the urban unemployment problem. Rural development is essential.”

Fields (1982) and Schultz (1982), like other migration researchers, incorporate the main insight of Harris and Todaro (1970) by including employment rates at both the origin and the destination as determinants of migration along with a host of other variables, finding confirmation of their importance. These articles also reflect the variety of empirical specifications used in migration research, with Fields estimating a linear regression explaining migration flows and Shultz estimating a multinomial logit model.

In addition to labor-market conditions, another important determinant of migration is distance, which is highlighted by gravity models. Intuition suggests that a greater spatial gap between two regions means a lower flow of migrants between them. Schwartz (1973) argues that a longer distance may raise the psychic cost of moving while decreasing the information available about the destination. The psychic cost arises because a longer migration distance means less frequent reunions with family and friends left behind. Even though the advancement of communication and media technologies weakens the foundation for these hypotheses, empirical evidence shows that distance plays an important role in individual migration decisions. The effect is seen in the studies of Fields (1982) and Schultz (1982) and virtually every other migration study.

21.2.2 More recent research

Some recent additions to the body of migration research draw on the past seminal studies while improving their empirical methods and exploiting relatively richer information in newer datasets. Other recent additions identify and study new factors in the migration decision.

21.2.2.1 Adding covariates to combat selection bias

Recent research recognizes that unobservable family and community characteristics of migrants can be different from the characteristics of those who stay behind, possibly biasing estimates of the impacts of observables on migration decisions. This selectivity effect may be different for distinct migrant destinations as well as for different sectors of employment at the destination. One way to reduce selectivity bias is to measure and include more household and community-level variables in the empirical models, thus reducing the number of omitted factors. Accordingly, household variables such as assets and the human capital of nonmigrant household members are allowed to have an impact on migration decisions via their effect on migration costs and remittances. Life cycle-related variables, such as marital status or changes in status, influence the migration decision as well. Added community variables include transportation access to commercial centers, in the belief that good access raises the incentive to migrate. “Network” variables capturing the presence of previous migrants at the destination, which can help migrating individuals adapt and find jobs, hence inducing them to migrate, may also have an influence.

Using nationally representative data on rural households in Mexico, [Mora and Taylor \(2007\)](#) estimate multinomial logit models that capture the effects of individual, family, and community variables on observed migration outcomes. The estimates show that greater family landholdings imply less migration and that better transportation links (measured by service frequency) encourage migration. The presence of nonfarm enterprises in the village makes migration to destinations with such jobs less likely. Furthermore, Mora and Taylor’s results indicate that the presence of other household members at a destination encourages migration to it. This network effect is also studied by [Giulietti et al. \(2014\)](#) using data from China. Building on the distinction between weak and strong ties in social-network theory ([Granovetter, 1973](#)), they distinguish between the presence at the destination of immediate family members (strong ties) and the presence of other residents from the same village (weak ties). A theoretical model predicts a larger migration effect from weak ties, and the Chinese evidence supports this prediction. Finally, [Marre \(2009\)](#) shows that family size and home ownership are important factors reducing the incentive to migrate as they are strongly and positively associated with the costs of moving.

21.2.2.2 Introducing access to public services in migration decisions

Another set of studies draws attention to availability of basic public services as a determinant of migration. With the goal of evaluating the importance of this push factor relative to the traditional pull factors, [Lall et al. \(2009\)](#) study migration from lagging to leading regions in Brazil. They combine a rich dataset of public services at the municipality level with individual records from four decades of Brazilian census data to evaluate the relative importance of wage differences and public services in the migrant’s decision to move. Predictably, wage differences are the main factor influencing migration choices. While basic public services are not important in the decision to move for better-off rural

residents, access to such services matters for the poor. Indeed, poor migrants are willing to accept lower wages to get access to better services. A Brazilian minimum-wage worker earning R\$7 per hour (about US\$2.30 in February 2008) was willing to pay R\$420 a year to have access to an additional hospital, R\$87 for a one percentage point increase in the likelihood of water access, and R\$42 for a one percentage point increase in the likelihood of electricity access. Although these results show that better public-service access benefits individual migrants, Lall, Timins, and Yu argue that the economy as a whole may end up worse off since the relocation is likely to add to congestion in urban areas without creating offsetting productivity benefits.⁵

These authors also raise an important methodological issue by pointing out that ignoring public service differentials may bias the estimated effects of wage differentials. The reason is that places with more job opportunities may also have better public services. By failing to control for public-service differentials, econometric estimates may thus overstate a migrant's willingness to move in response to wage differences.

The effect of the public-services push factor is seen in Sri Lanka ([World Bank, 2010](#)) where lack of access to basic public services such as water and electricity also influences migration decisions. Using data from the Sri Lanka Integrated Survey for working-age people between the ages of 15 and 49 years, the analysis finds that migration decisions in the 1990s were influenced by district-level differences in access to well water and electricity, particularly for the less educated. For individuals with a secondary-school education or less, a 1% difference in the share of well-water coverage between the origin and the destination increased the likelihood of moving by 0.5%. Water supply differentials did not matter for the better educated.

Recent evidence from Nepal also points to the importance of access to basic services in influencing migration decisions. [Dudwick et al. \(2011\)](#) use the population census in 2001 with the objective of understanding why migrants are attracted to particular locations. They examine the roles of income, access to basic services, and physical and social distances in influencing rural–urban migration decisions. The indicators of basic services include access to facilities such as schools, hospitals, markets, and banks. While they find the usual effects of distance on migration, the results show that migrants choose destinations with better access to schools, hospitals, and markets. Moreover, a terrain elevation variable, meant to capture transport barriers, has a negative effect.

From a policy perspective, all of these findings imply that providing access to basic services can help eliminate a major push factor that leads to migration from rural areas in developing countries.⁶ Development policies should recognize the importance of such

⁵ See [Ferre \(2009\)](#) on the impact of internal migration for the receiving urban areas in Brazil.

⁶ Economies of scale in the provision of health services and other public services would lead to a counter-argument in favor of spatial concentration of these services, with dispersion across both urban and rural areas desirable on equity grounds but potentially inefficient.

access as a determinant of the population distribution in these countries. Mourmouras and Rangazas (2013) offer a theoretical analysis of this issue, analyzing a model with rural–urban migration where the government allocates public services between the two areas with the goal of maximizing the country’s overall welfare.

21.2.2.3 *Climate migrants*

In addition to the influence of public services on migration, other important drivers include environmental factors. According to the Intergovernmental Panel on Climate Change, climate change will degrade the environment considerably during this century (Reuveny, 2007),⁷ with the impacts felt more in developing countries than in developed countries, because of both geography and a more limited scope for policy intervention. Thus, environmental conditions may play an increasingly important role as a push factor as they threaten the livelihoods of rural residents. For this reason, the environmental dimension of migration has recently begun to gain the attention of researchers.

From this perspective, scholars view migration as an adaptation strategy in the face of worsening environmental conditions due to drought, soil quality deterioration, and deforestation.⁸ This focus aligns with a new perception of migration among labor economists as a household strategy for income diversification in response to environmental and other risks. Hunter et al. (2011), for example, study the effect of rainfall on emigration from rural Mexico. The study focuses on international migration (specifically to the United States), but the same force affects internal migration within both Mexico and other developing countries. The results indicate an association between rainfall patterns and migration, where dry years cause a migration push and wet years inhibit migration from rural areas.

Barrios et al. (2006) investigate the effect of rainfall on urbanization in sub-Saharan African. They point out that the sub-Saharan African economies are particularly dependent on rainfall, which has been declining since the late 1950s. This dependence makes the agriculture-dependent rural population highly vulnerable to variations in rainfall, potentially affecting rural–urban migration patterns. They assert that “climate change scenarios tend to suggest that extreme climate variations are likely to cause abrupt changes in human settlements and urbanization patterns in sub-Saharan Africa more than anywhere else in the world.”

Reuveny (2007) also provides a careful analysis of climate change-induced migration across countries, providing information on environmental push factors at individual country levels. He states that 20–30 million people moved from Gansu and Ningxia

⁷ In particular, Reuveny (2007) reports that since the 1950s, the average global temperature rose by about 0.1 °C per decade, winter snow cover declined by 10%, the frequencies and intensities of droughts, storms, and warm periods rose, and the sea level rose by 20 cm.

⁸ See Henry et al. (2003) for a study of the effect of environmental factors on internal migration in Burkina Faso.

provinces to urban centers in China during the 1980s and 1990s because of floods, land degradation, desertification, and water scarcity. Similarly, Reuveny's evidence shows that 600,000–900,000 people migrated from the rural areas of Mexico to urban centers and the United States because of environmental degradation combined with other factors. Finally, around 70,000 people moved from the Arctic region in Russia to urban centers because of extreme weather conditions and socioeconomic decline in the 1990s.⁹

21.2.2.4 Income risk as a factor governing migration

The income risk faced by migrants at their destination has been recognized as a factor in the migration decision, and it is partly addressed by studies that measure the presence of other household members at the destination, which may ease the job-finding process. However, several recent studies model the sources of income risk in a more fundamental fashion while carrying out empirical tests. The studies of [Bryan et al. \(2014\)](#) and [Munshi and Rosenzweig \(2013\)](#) both note the existence of unexploited migration opportunities in particular developing countries, with the first article investigating a famine-prone region of Bangladesh where migration presumably has large benefits but is puzzlingly low, and the second article focusing on India, where rural–urban migration is low relative to that in other countries despite a large rural–urban wage gap that would appear to make it attractive. Both studies attribute the lack of migration to the income risk faced by migrants, providing different types of evidence in favor of this view.

[Bryan et al. \(2014\)](#) hypothesize that a migrant does not know in advance his suitability for urban employment (e.g., whether he will be trusted by a rickshaw owner to operate this valuable asset as a driver), which makes incurring the cost of migration risky. The study carried out an experiment where some residents of the famine-prone Bangladesh region of Rangpur were offered a subsidy to cover migration costs if they chose to temporarily migrate to an urban area during the preharvest famine months. The subsidy predictably increased migration, but the striking observation was an increase in migration by the same individuals in the subsequent year, when the subsidy was not offered. This observation is consistent with a reduction in income risk as subsidized migrants learned their suitability for city jobs, acting on this knowledge in the next season. Bryan et al. developed a theoretical model capturing this phenomenon, which also showed that the households closest to subsistence should be least prone to migrate in the absence of a subsidy, fearing a catastrophic drop in disposable income in the event of costly, unfruitful migration. This prediction also finds empirical support in the article.

[Munshi and Rosenzweig \(2013\)](#) argue that migration income risk in India comes from the loss of support from the local caste-based insurance network in the rural home region. These networks provide a form of income pooling designed to help households suffering drops in income, and migration to the city reduces or eliminates the support that

⁹ For further details, see Table 1, panel B in [Reuveny \(2007\)](#).

can be expected from the network. As a result, income risk is higher in the city than in the home region, but for a reason different from that in [Bryan et al. \(2014\)](#). The resulting empirical prediction, supported by a theoretical model, is that higher-income rural households, which tend to subsidize rural households through the network rather than benefiting from it, should be more likely to migrate than lower-income households. The basic empirical results in the article confirm this prediction, with additional effort expended on estimation and simulation of a structural model.

21.3. MODELS OF MIGRATION AND CITY SIZES IN DEVELOPING COUNTRIES

As seen in [Section 21.2](#), rural–urban migration is a major force that drives city growth in developing countries. This same migration process led to the high urbanization levels that now exist in developed countries, but the process was largely complete long ago. Given the centrality of rural–urban migration in developing countries, researchers have offered various formal treatments showing how the migration process determines city sizes. These models formalize the incentives for migration and derive the condition that characterizes a migration equilibrium, where the population flow from the countryside stops.

As recognized in all of the empirical work discussed in [Section 21.2](#), the incentive for migration in the theoretical models depends on the difference in living standards between the rural and urban areas. When the urban standard of living exceeds the rural standard, an incentive exists for rural–urban migration, so that population leaves the countryside and cities grow. Migration equilibrium is achieved when urban and rural living standards are equalized. The costs of migration are disregarded, being implicitly assumed to equal zero.

The models are mostly silent about the forces that produce a surge of rural–urban migration from a starting point where migration is absent. But the implicit scenario is one where modern production technologies are adopted by urban firms, greatly increasing worker productivity and wages and thus creating a large gap between urban and rural living standards that prompts migration to the city.

To reach a migration equilibrium, the rural–urban population flow must close this gap between living standards, and the models differ in their portrayals of the forces that achieve this closure. As noted above, the seminal article of [Harris and Todaro \(1970\)](#) assumes that the equilibrating force is adjustment of the urban unemployment rate. This rate rises as migrants flow into the city, reducing a migrant’s chance of finding a well-paying job and ultimately choking off migration. Another equilibrating force not captured in the Harris–Todaro model is escalation of the urban cost of living as migrant flows raise the city’s population. This cost-of-living increase will arise mainly in the housing market through rising rents, a force that is captured in the model of [Brueckner \(1990\)](#). Once living costs have risen enough to offset the benefit of a higher urban

income, the incentive for migration is eliminated. The hybrid models developed by Brueckner and Zenou (1999) and Brueckner and Kim (2001) combine these approaches by assuming that increases in both the urban cost of living and unemployment jointly serve to equilibrate the migration process.

Since all of these models are static in nature, predicting the achievement of a stable population split between rural and urban areas, they do not match the reality in developing countries, where rural–urban migration is ongoing. This mismatch can be remedied by embedding the models in a dynamic context, where the rural and urban populations slowly adjust toward the equilibrium predicted by the static model. Alternatively, the reality could involve continual disruption of static equilibria (resulting, for example, from technical change that leads to an upward time trend in urban incomes), with the economy constantly on the move to new equilibria with larger city sizes.

A further question concerns the economic efficiency of the migration equilibria achieved under the models. The efficiency question is whether rural–urban migration produces the “right” division of the population between the city and the countryside from society’s point of view. The following subsections discuss the models and consider this question.

21.3.1 The Harris–Todaro model

Many variables are potentially endogenous in a model of rural–urban migration. In addition to the rural and urban populations themselves, these variables include the prices of the (nonhousing) goods produced in the city and countryside, the wages of workers in these two locations (which depend partly on the output prices), and the capital and land inputs used along with labor in urban and rural production. Harris and Todaro (1970) assume that the capital and land inputs are fixed, but they allow wages and output prices to adjust with the rural–urban population split. However, since nonhousing price adjustment is largely inessential in describing the main elements of the models of rural–urban migration, it will be suppressed in the ensuing discussion. Instead, the prices of the nonhousing goods produced in the rural and urban areas will be fixed, reflecting the assumption that the goods produced in both locations are traded on world markets. In other words, portions of both the agricultural good produced in the countryside and the manufacturing good produced in the city are exported, so local prices must reflect world prices, which are fixed from the point of view of any one country.

With the fixed output prices both set equal to 1, annual wages are then just equal to the (annual) marginal product of labor, which diminishes with the number of workers employed. Denoting the city population by N , the urban income level is then given by $\gamma_u(N)$ and the rural income is given by $\gamma_r(\bar{N} - N)$, where $\bar{N}$ is the fixed total population of the country. The functions γ_u and γ_r , which give the (annual) marginal products of labor in the two locations, are constant or decreasing in the usual fashion

($y'_u, y'_r \leq 0$). Note that, while the analysis assumes the existence of a single city containing the entire urban population, this assumption can be relaxed to allow for multiple cities.¹⁰

The Harris–Todaro model suppresses decreasing housing (or land) consumption, with workers in both the city and the countryside implicitly assumed to consume only the outputs of the urban and rural production sectors. With consumption prices thus being the same regardless of a worker's location, urban and rural costs of living are identical, and income differences alone guide migration decisions. In the absence of any restrictions on the urban labor market, rural residents would want to move to the city as long as $y_u(N) > y_r(\bar{N} - N)$. The condition for migration equilibrium would then be

$$y_u(N) = y_r(\bar{N} - N), \quad (21.1)$$

which determines the city population size that equates urban and rural incomes. If marginal productivity tends to be higher in city employment than in the rural sector (with $y_u > y_r$ holding at a common population size), substantial migration would be required to equate urban and rural incomes, with N possibly approaching $\bar{N}$.

The Harris–Todaro model modifies the equilibrium condition in (21.1) to incorporate another feature of cities in developing countries that has been claimed to coexist, in a puzzling fashion, with large rural–urban population flows: high urban unemployment. To explain the apparent anomaly of rural migrants moving to the city despite a large chance of being unemployed, Harris and Todaro observed that the *expected* urban income (in a probabilistic sense) can still be high relative to the rural income if wages while employed in the city are high.

To formalize this idea, Harris and Todaro assumed the existence of an urban minimum wage, yielding annual income of $\bar{w}$. At this wage, urban employers are willing to hire J workers, where J satisfies $y_u(J) = \bar{w}$. With urban jobs thus fixed at J , migration to the city has no effect on the incomes of those employed, although it does reduce the expected wage. Assuming the absence of any unemployment assistance, the expected urban income equals the probability of employment times $\bar{w}$, or $(J/N)\bar{w}$. The migration equilibrium condition now equates this expected income value to the (certain) rural income level, being written as

$$\frac{J}{N}\bar{w} = y_r(\bar{N} - N), \quad (21.2)$$

a condition that assumes risk neutrality on the part of migrants. Note that while an increase in N previously reduced the urban marginal product in (21.1), helping to decrease the attractiveness of further migration, a higher N now reduces the chance

¹⁰ With m cities of population N , the rural population would be $\bar{N} - mN$. To endogenize the number of cities, an approach like that of Henderson and Wang (2005) could be used.

of urban employment, with the same equilibrating effect (which is reinforced by the migration-induced rise in γ_r).

If the rural marginal product of labor is constant, making rural income constant at $\bar{y}_r$, then (21.2) gives a direct solution for N :

$$N = \frac{\bar{w}}{\bar{y}_r} J, \quad (21.3)$$

which indicates that the urban population equals a multiple of the number of available jobs ($\bar{w}/\bar{y}_r > 1$ holds). From (21.3), an increase in $\bar{y}_r$ reduces N by raising the attractiveness of the countryside, while an increase in J with $\bar{w}$ held fixed (caused, say, by an increase in urban productivity) raises the chance of employment and leads to an offsetting increase in N that restores the original employment probability. Since an increase in $\bar{w}$ also leads to a decrease in J , the effect on N is ambiguous, with the sign depending on the elasticity of the γ_u function. N rises if a higher $\bar{w}$ raises $\bar{w}J$, falling otherwise.

It should be noted that because prices are fixed in this version of the Harris–Todaro model, the paradox described in Section 21.3, where an increase in J leads to an increase in the unemployment rate (a decrease in J/N) cannot occur. This outcome can be seen directly in (21.3), where J/N is constant, and it also can be established when γ_r is variable by differentiating (21.2). Thus, price flexibility is required for the paradox to emerge.

21.3.2 Urban cost of living as an equilibration mechanism

As explained above, the cost of living is the same in rural and urban areas under the Harris–Todaro model, ruling it out as an equilibrating force. To introduce cost-of-living differences, it is natural to use the standard urban model developed by Alonso (1964), Mills (1967), and Muth (1969) to capture the effect of population size on a city's cost of living, following Brueckner (1990). In the standard model, all residents commute to jobs at the city center, paying an (annual) transportation cost of t per mile, so tx gives the cost of commuting from a residential location at distance x from the center. Urban residents consume housing, represented as direct consumption of land in the simplest form of the model, along with a composite nonhousing good made up of the outputs of the urban and rural sectors. Residents living far from the center, who incur high commuting costs, are compensated with lower land rents. Land rent p thus declines as distance x from the center increases, with rent falling to the fixed rural rent p_r at the edge of the city, which lies at distance $\bar{x}$.

The higher urban cost of living relative to that in rural areas is due to a combination of higher land rents and commuting-cost outlays. By comparison, rural workers spend nothing on commuting by living near the fields they till, and (like farmers) they pay p_r for the land they consume, an amount less than urban residents pay in the city's interior. But since an urban resident living at the edge of the city pays the same land rent as rural residents (p_r), his higher cost of living comes entirely in the form of a higher commuting

cost. After paying this cost, the disposable income of the edge resident is equal to $y_u - t\bar{x}$. Since they face the same land rent, rural residents will therefore reach the same utility level as the city's edge resident if their income y_r is equal to $y_u - t\bar{x}$. But since the utility levels of all city dwellers are equal in the urban equilibrium (matching the edge's resident's utility), the condition $y_u - t\bar{x} = y_r$ ensures that the rural utility will equal that of *an urban resident living anywhere in the city*. This condition, therefore, becomes the migration equilibrium condition when both incomes and the urban cost of living are taken into account.

To use this condition to analyze the properties of the migration equilibrium, comparative-static results from the standard urban model can be exploited. [Wheaton \(1974\)](#) first presented these results, showing that $\bar{x}$ is an increasing function of the urban population N , a natural conclusion. He also established that $\bar{x}$ increases with the urban income level y_u (richer cities take up more space) and that $\bar{x}$ decreases with t , the commuting cost per mile (cities with high commuting costs are more compact). Therefore, $\bar{x}$ can be written as $\bar{x}(N, y_u, t)$, with $\bar{x}_N > 0$, $\bar{x}_{y_u} > 0$, and $\bar{x}_t < 0$, where the subscripts denote partial derivatives.¹¹

Recognizing that the income levels generally depend on N , we can write the migration equilibrium condition as

$$y_u(N) - t\bar{x}(N, y_u(N), t) = y_r(\bar{N} - N). \quad (21.4)$$

The value of N that satisfies this condition gives the equilibrium size of the urban population. To explore the implications of this condition, it is helpful to follow [Brueckner \(1990\)](#) by initially assuming constant marginal products of labor and thus fixed urban and rural incomes. Substituting $\bar{y}_u$ and $\bar{y}_r$ in place of the income expressions in (21.4), the condition becomes

$$\bar{y}_u - t\bar{x}(N, \bar{y}_u, t) = \bar{y}_r, \quad (21.5)$$

which determines the urban population N as a function of the remaining variables, $\bar{y}_u$, $\bar{y}_r$, and t .

From the signs of the $\bar{x}$ derivatives, comparative-static analysis of the equilibrium is simple and intuitive. When $\bar{y}_r$ increases, the left-hand side of (21.5) must increase as well, which requires a decline in $\bar{x}$ and thus a decline in N . Formally,

$$\frac{\partial N}{\partial \bar{y}_r} = -\frac{1}{\bar{x}_N} < 0. \quad (21.6)$$

Thus, as higher rural income makes the countryside more attractive, the urban population falls. This population decline reduces the urban cost of living (as represented by

¹¹ Another result is that $\bar{x}$ decreases with the rural rent level, but this conclusion is not needed in the present context.

the commuting cost of the edge resident), making the city and countryside equally attractive again.

Additional results rely on further conclusions of Wheaton (1974), who showed that the urban utility level rises with the urban income level $\bar{y}_u$ and falls with the commuting cost per mile t , results that are natural. With the land price paid by the edge resident fixed at p_r , these utility changes must be mirrored in the change in disposable income for the edge resident, which therefore must rise with y_u and fall with t . From differentiation of the left-hand side of (21.5), $1 - t\bar{x}_{y_u} > 0$ and $-\bar{x} - t\bar{x}_t < 0$ must then hold even though the signs are not clear from inspection. Totally differentiating (21.5) then yields

$$\frac{\partial N}{\partial \bar{y}_u} = \frac{1 - t\bar{x}_{y_u}}{\bar{x}_N} > 0, \quad \frac{\partial N}{\partial t} = -\frac{\bar{x} + t\bar{x}_t}{\bar{x}_N} < 0. \quad (21.7)$$

Intuitively, when y_u rises, the city becomes more attractive, and its population grows until the cost of living (represented by the commuting cost for the edge resident) has risen enough to nullify the gain. Similarly, when t rises, indicating the presence of an inferior transport network, the city becomes less attractive as the commuting cost of the edge resident rises. The population must then fall to reduce the edge resident's cost to its original level.

Returning to the general form (21.4) of the equilibrium condition, where incomes are no longer fixed, only one parameter remains—namely, t . An increase in t reduces the left-hand side of (21.4) as before, reducing the difference between the expressions on the left-hand side and the right-hand side, but the required change in N (which must raise this difference in an offsetting fashion) is not immediately clear. However, using previous results, the derivative of the difference between the left-hand side and right-hand side of (21.4) with respect to N is $-t\bar{x}_N + [1 - t\bar{x}_{y_u}]y'_u + y'_r < 0$. Therefore, a decline in N is required to raise the difference, offsetting the higher t . The inverse relationship between N and t is thus preserved under the general model, and the comparative-static effects for income can also be recovered by introducing shift factors that move the y_u and y_r functions. Writing these functions as $y_u(N) = \alpha_u f_u(N)$ and $y_r(N) = \alpha_r f_r(\bar{N} - N)$, it is easily seen that an increase in α_u raises N , while an increase in α_r has the opposite effect. Thus, an increase in urban productivity encourages additional migration to the city, while an increase in rural productivity prompts a return to the countryside, as before.

21.3.3 Hybrid models

Brueckner and Zenou (1999) and Brueckner and Kim (2001) develop hybrid models where unemployment and cost-of-living adjustments combine to equilibrate rural–urban migration. In Brueckner and Zenou's model, employed and unemployed residents constitute distinct groups who live in different parts of the city, with the unemployed residents subsisting on government-assistance payments. Brueckner and Kim offer a simpler model where employment status is determined anew each period, with workers

also smoothing their incomes via saving as they alternate between employment and joblessness. With this approach, the city can be modeled as containing a single group of residents with incomes equal to the expected value of income across the employed and unemployed states. Assuming as before that assistance payments are zero, this expected income is again $(J/N)\bar{w} \equiv w_e$.

Although Brueckner and Kim carry out their analysis from first principles, an equivalent and simpler approach relies on the framework from above. In particular, the migration equilibrium condition comes from simply replacing $y_u(N)$ in (21.4) with w_e , so equilibrium is characterized by the following two conditions:

$$w_e - t\bar{x}(N, w_e, t) = y_r(\bar{N} - N), \quad (21.8)$$

$$w_e = \frac{J}{N}\bar{w}. \quad (21.9)$$

Since w_e decreases with N like $y_u(N)$, this hybrid model has the same properties as the general model analyzed above. In particular, N falls with an increase in t or with an increase in a rural productivity parameter α_r . In addition, N increases with J , a change that would result from an increase in an urban productivity parameter α_u . As before, the effect of an increase in $\bar{w}$ depends on the elasticity of the y_u function. N again rises if a higher $\bar{w}$ raises $J\bar{w}$, falling otherwise.

It is easy to see that the Harris–Todaro paradox cannot arise in this hybrid model, as in the basic model above. In particular, N must rise by less than any increase in J , reducing the unemployment rate. To see this conclusion, note that if N were to rise enough to keep J/N constant, the urban standard of living would nevertheless fall given the escalation in land costs. With the decline in the rural population causing y_r to rise or stay constant, it follows that the standard of living is lower in the city, implying that N has increased too much. As a result, J/N must be higher in the new equilibrium.

The stability of the various equilibria considered so far has not been considered. Stability requires that the urban disposable income is greater than (less than) the rural income when N is below (above) its equilibrium value. When this condition is satisfied, migration incentives cause N to rise when it is below the equilibrium value and fall otherwise, yielding convergence to the equilibrium. The stability condition implies that left-hand-side expression minus the right-hand-side expression in an equilibrium condition must be a decreasing function of N . It is easy to see that this condition is satisfied for each of the equilibria considered above.

21.3.4 Dynamics

The previous models are static in nature, generating a city of equilibrium size and an absence of rural–urban migration. The reality in developing countries, however, is decades of ongoing migration with no apparent end in sight. One way of reconciling this

reality with the previous models is to imagine that the adjustment to equilibrium is sluggish (see Brueckner, 1990). To formalize this idea, use the hybrid model, and let $N(t, \alpha_r, J\bar{w})$ denote the equilibrium urban population as a function of the model parameters. Let τ denote time, and assume that the parameters are independent of τ . Then, suppose that the population evolves according to a partial-adjustment process given by

$$\Delta N_{\tau+1} \equiv N_{\tau+1} - N_{\tau} = \lambda [N(t, \alpha_r, J\bar{w}) - N_{\tau}], \quad (21.10)$$

where λ is an adjustment parameter satisfying $0 < \lambda \leq 1$. Thus, a fraction λ of the gap between the equilibrium and the current urban population is closed each period, leading to a potentially long adjustment process. Starting with a small city size, a large equilibrium N (a result of a low t or α_r or a high $J\bar{w}$) would lead to rapid growth in the urban population, while growth would be slower with an equilibrium N of more moderate size.

This approach could be modified by allowing t , α_r , and $J\bar{w}$ to be time dependent. Then, the equilibrium population would be a moving target, and ongoing changes in N would be due to both the changing equilibrium and the sluggishness of adjustments.

Taken literally, a static model should imply instantaneous adjustment to the equilibrium urban–rural division of the population, with no ongoing migration. The dynamics in (21.10) are therefore ad hoc in nature, with a different type of model required to properly generate continuing migration. Lucas (2004) develops such a model, assuming that urban residents devote some of their time to accumulation of human capital. When this structure is combined with the assumption of human capital externalities, where individual capital accumulation is faster the higher is the capital level of the city’s most skilled residents, the model generates a gradual emptying of the countryside (the details are complex). Henderson and Wang (2005) offer a simpler, related model in which savings (which occur at a fixed rate) are devoted to human-capital accumulation, with this capital raising individual productivity more in cities than in the rural area. This growing relative urban income advantage, combined with the need to equate rural and urban living standards, requires an ongoing population shift from the countryside to cities, matching the spirit of the models analyzed above.¹²

21.3.5 Socially optimal city sizes

Is the city size generated by the rural–urban migration equilibrium efficient, maximizing social welfare? To answer this question in the present context, the natural welfare function is the value of the economy’s output minus resource costs. Since capital inputs are fixed in the previous models, capital costs can be disregarded in characterizing the social optimum, with the only resource cost being the commuting cost incurred by urban

¹² Their setup is actually a bit more complicated than described, since city sizes are chosen by profit-maximizing land developers and wages depend on the endogenous output price of the city-produced good. However, the underlying mechanism appears to work as described.

residents. Aggregate commuting cost is given by $ACC \equiv \int_0^{\bar{x}} [2\pi x/q(x, N)] t x dx$, where $q(x, N)$ is land consumption at distance x from the urban center and $1/q(x, N)$ equals population density. N is an argument of q because individual land consumption is in general a decreasing function of the city's population, as shown by Wheaton (1974).¹³ The integral weights population density by the land area of the ring of land at distance x (equal to $2\pi x dx$) to get the ring population and then multiplies it by tx , the commuting cost of each ring resident, with the result summed across all locations in the city.

Consider first the simple case where the marginal products of urban and rural labor are constant and given by $\bar{y}_u$ and $\bar{y}_r$. Also, suppose that individual land consumption is fixed at one unit of land, thus being independent of x and N . Evaluating the above integral in this case yields $ACC = 2t\pi\bar{x}^3/3$. Since the urban and rural outputs are just $N\bar{y}_u$ and $(\bar{N} - N)\bar{y}_r$, the value of the economy's output net of commuting cost is

$$N\bar{y}_u + (\bar{N} - N)\bar{y}_r - 2t\pi\bar{x}(N)^3/3, \quad (21.11)$$

where the N argument of $\bar{x}$ is reintroduced (the previous income and t arguments are unneeded and thus suppressed). The socially optimal N maximizes this expression. To carry out the maximization, the relationship $N = \pi\bar{x}^2$ must be used, which says that the total population equals the city's land area (a consequence of $q = 1$). Differentiating this expression yields $\bar{x}_N = \partial\bar{x}/\partial N = 1/2\pi\bar{x}(N)$. Then, the socially optimal N , which comes from differentiating (21.11) and setting the result equal to zero, satisfies

$$\bar{y}_u - \bar{y}_r - 2t\pi\bar{x}(N)^2 \frac{\partial\bar{x}}{\partial N} = \bar{y}_u - t\bar{x}(N) - \bar{y}_r = 0. \quad (21.12)$$

Since this condition is the same as (21.5), the migration equilibrium condition when incomes are fixed, the equilibrium in this case is efficient. Note that the stability condition for the equilibrium (the last expression in (21.12) decreases with N) means that the second-order condition for the social optimality problem is satisfied.

Efficiency is obtained because no externalities are present under the previous assumptions. One assumption was fixed land consumption, and when this assumption is relaxed, an increase in N generates an externality by making the city denser as q falls in response to the higher population. By putting people closer to the center on average, rising density saves commuting costs, so that the higher N leads to a positive externality. Since this externality is not taken into account by migrants, the equilibrium city population is too small. Letting $\Omega(N)$ capture the positive density externality (derived in the Appendix), the optimality condition is¹⁴

¹³ Actually, land consumption q depends positively on the urban utility level u , but since the equilibrium u is a decreasing function of N , q decreases with N .

¹⁴ The second-order condition for this problem is assumed to hold (it cannot be checked because of the complexity of the derivative $\Omega'(N)$).

$$\bar{y}_u - t\bar{x}(N) + \Omega(N) = \bar{y}_r. \quad (21.13)$$

This condition implies that $\bar{y}_u - t\bar{x}(N) < \bar{y}_r$ holds at the optimum, implying that $\bar{x}$ and hence N are larger at the optimum than in the equilibrium. Thus, the density externality that affects commuting distances tends to make cities too small.

Another population externality operating through commuting costs, which is negative in direction, would arise if the model were to include traffic congestion. Then, while an increase in N would make the city denser and reduce the average commuting distance, the higher traffic volumes from a larger population would worsen congestion, raising commuting costs. This effect could dominate, making the net population/commuting-cost externality negative. While the model would have to be restructured to embrace traffic congestion, the resulting framework would probably imply that N is larger, not smaller, at the optimum than in the equilibrium.

Externalities can also arise in production. To begin consideration of this case, suppose that incomes are no longer constant, but that land consumption is again fixed at unity. Let $Y_u(N)$ and $Y_r(\bar{N} - N)$ be the total urban and rural outputs, with marginal products equal to $y_i(N) \equiv Y'_i(N)$, $i = u, r$.¹⁵ The sum of Y_u and Y_r minus ACC is the objective function, and following (21.12), the first-order condition is $y_u(N) - t\bar{x}(N) = y_r(\bar{N} - N)$. This condition coincides with the equilibrium condition (21.4) in the variable-income case, indicating efficiency of the equilibrium, and the reason is that a production externality is not yet present.¹⁶

To introduce such an externality, let production in the city exhibit external economies of scale, whose source is urban agglomeration economies. In this case, the urban production function is $A(N)Y_u(N)$. The shift factor $A(N)$, which is viewed as parametric by individual firms, satisfies $A'(N) > 0$ at low values of N , indicating economies of scale, but $A'(N)$ could become negative at large values of N , reflecting eventual diseconomies ($A'' < 0$ is assumed). In this setting, the marginal product of labor from the perspective of a firm (which gives the urban income level) is $A(N)Y'_u(N) = A(N)y_u(N)$.

¹⁵ At this point, it is worth noting an interpretational issue that arises when assuming that marginal products equal consumer incomes while also being functions of the population. In effect, the model implicitly portrays the city as containing a single firm with production function $Y_u(N)$ while assuming that, despite its size, the firm behaves competitively in the labor market, paying a wage equal to the marginal product rather than exploiting the upward-sloping supply curve of labor faced by the city. This issue, which is present in the article by [Harris and Todaro \(1970\)](#) and elsewhere in the literature, does not arise when labor's marginal product is constant. In that case, the city's production can be viewed as coming from a large number of small, competitive firms, each paying a wage equal to the marginal product in the usual fashion.

¹⁶ With land consumption fixed, the dependence of $\bar{x}$ on income disappears. If land consumption were instead variable, the impact of N on $\bar{x}$ via $y_u(N)$ would need to be considered in the social optimization problem.

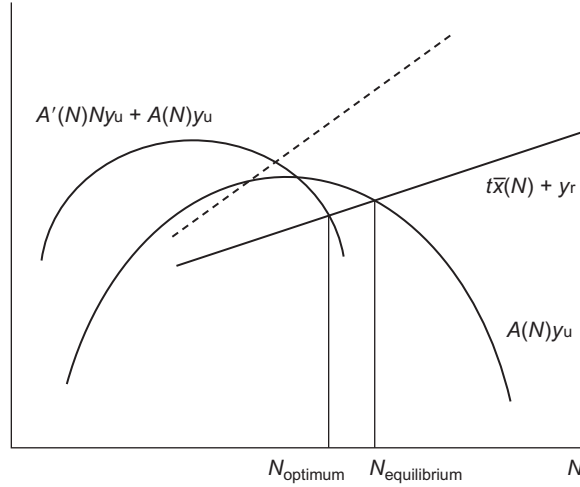


Figure 21.1 Equilibrium and optimum.

The socially optimal N maximizes $A(N)Y_u(N) + Y_r(\bar{N} - N) - ACC$, and the derivative of this expression with respect to N is

$$A'(N)Y_u(N) + A(N)\gamma_u(N) - t\bar{x}(N) - \gamma_r(\bar{N} - N). \quad (21.14)$$

If (21.14) decreases everywhere with N , the optimum is found by setting the expression equal to zero, as in the previous cases (the second-order condition then holds). If $A'(N)$ is positive at the optimum, then the migration equilibrium condition, which is $A(N)\gamma_u(N) - t\bar{x}(N) = \gamma_r(N)$, yields a smaller value of N ((21.14) includes an extra positive term). As a result, the equilibrium city size is too small, with migrants ignoring the external productivity benefits they generate in moving to the city. If, however, $A'(N) < 0$ holds at the optimum, with the optimal city so large that external diseconomies have begun to set in, then the conclusion is reversed: the equilibrium city size is too large, with migrants ignoring their negative external effects.

This outcome is illustrated in Figure 21.1, which assumes for simplicity that γ_u and γ_r are constants, independent of N , with $Y_u(N) = N\gamma_u$. $A(N)$ is an inverted U-shaped function, and the equilibrium N lies at the intersection of $A(N)\gamma_u$ (also U shaped) and the upward-sloping line corresponding to $t\bar{x}(N) + \gamma_r$, as shown in Figure 21.1.¹⁷ The optimum (from (21.14)) lies at the intersection of $A'(N)N\gamma_u + A(N)\gamma_u$ and the same line, yielding a lower value of N , as shown. Even when $A(N)$ is U shaped, the equilibrium could lie in a range where the function is upward sloping, yielding the reverse

¹⁷ This expression generally need not yield a linear relationship, but one is drawn for convenience. Note also that the other intersections of the line with the U-shaped curves (which are not shown) are not relevant, yielding an unstable equilibrium or welfare minimum.

relationship between the equilibrium and the optimum (with the optimal N larger). This outcome occurs when $t\bar{x}(N) + \gamma_r$ corresponds to the dotted line rather than solid line in Figure 21.1. Note that with γ_u constant, the divergence between the equilibrium and the optimum can be seen as arising from a migrant's consideration of his average product in the city ($A(N)\gamma_u$) rather than his marginal product ($A'(N)N\gamma_u + A(N)\gamma_u$) in deciding whether to migrate.

In summary, the foregoing analysis shows that externalities can make cities either too large or too small in equilibrium relative to the socially optimal size. External diseconomies in production or the external effect of population on traffic congestion (which is not formally captured in the model) tend to make the city too large, whereas external economies in production or the population's external effect on densities (and thus commuting distances) tend to make the city too small. The lesson is that when externalities are present, the decentralized economy cannot be trusted to generate an optimal division of the population between the city and the countryside.

Such inefficiencies are also the focus of a body of literature in public economics initiated by [Flatters et al. \(1974\)](#), which studies the allocation of residents among urban jurisdictions providing local public goods. While the production-based externality arising via the $A(N)$ function is present in such models, an additional externality arises from the sharing of public-good costs. In particular, while an entering migrant makes his decision on the basis of the average (per capita) cost of a jurisdiction's public good, his entry drives down this average through cost sharing, an effect that he ignores. In such models, inter-jurisdictional transfers are needed to generate an efficient outcome, and such transfers can also achieve efficiency in the previous rural–urban context.

[Arnott \(1979\)](#) and [Au and Henderson \(2006\)](#) also analyze the optimal city size in a spatial economy, but they consider the optimal size of a city in isolation, not focusing on the optimal population split between the city and the countryside.¹⁸ In other words, they choose N to maximize $A(N)Y_u(N) - ACC$, without considering rural output. In addition to generating analytical results, [Au and Henderson \(2006\)](#) also ask whether Chinese cities are optimal in size according to this criterion. They estimate the function $A(N)Y_u(N) - ACC$ using Chinese data and then find the locations of existing cities relative to the function's maximum, concluding that city populations tend to be too small. The reason, they argue, is China's institutional restrictions on rural–urban migration (the hukou rules).

21.3.6 Empirical evidence

Although there is a vast empirical literature on various aspects of urbanization, the part of this literature that focuses on the determination of city sizes is most relevant to the

¹⁸ These models thus diverge from those in the public-economics tradition, which do not consider a single jurisdiction in isolation.

theoretical models discussed above. The regressions in Brueckner (1990) bear the closest connection since they are directly motivated by the model presented in Section 21.3.2. Using a small cross section of countries, the regressions relate measures of the extent of a country's urbanization to the urban–rural income ratio (y_u/y_r) and other variables, usually finding that this ratio has the expected positive effect (recall (21.6) and (21.7)). Additional studies, including those of Rosen and Resnick (1978), Wheaton and Shishido (1981), Ales and Glaeser (1995), Mutlu (1989), Moomaw and Shatter (1996), Davis and Henderson (2003), Henderson and Wang (2007), and Barrios et al. (2006), follow a similar approach by relating urbanization measures, sometimes including city growth or a measure of primacy (the extent of population concentration in the country's largest city),¹⁹ to a broader set of explanatory variables in cross-sectional or panel regressions. The variables that appear in at least one of these studies include the country's level of economic development (GDP per capita), sectoral employment shares (agriculture and manufacturing), administrative centralization (captured by a federal structure), a centrally planned economy, openness to international trade, land area, overall population, education and literacy, income inequality, ethnic heterogeneity, the extent of transportation linkages, measures of political freedom, and foreign-aid assistance. Higher GDP per capita reliably spurs urbanization, as does a federal political structure.

The study of Barrios et al. (2006), mentioned in Section 21.2, includes some of these same variables, but it is distinguished by a focus on low rainfall, which depresses rural incomes, as a factor spurring rural–urban migration and city growth (an effect that is measured using panel data for countries in Africa). Poelhekke (2011) offers a related analysis that focuses on rural income risk (possibly from rainfall variability) as a motivation for migration to cities. He shows that the growth of urbanized populations in a large international panel dataset responds positively to a time-varying, country-level measure of the volatility of agricultural value added, as predicted. In another novel study discussed in more detail in the next section, Hidalgo et al. (2010) focus on the migration of urban squatters, who enter the city through a “land invasion.” The study shows that land invasions in Brazil are more likely to occur and be larger when agricultural yields in the countryside are lower, indicating low rural incomes.

Young (2013) offers empirical evidence supporting a view at variance with the models described above. In particular, Young documents evidence showing a gap between living standards in urban and rural areas, rather than the equality that characterizes migration equilibrium in the preceding analysis. He argues that this gap reflects

¹⁹ Note that, as structured, the models considered previously are not capable of handling the phenomenon of primacy, where the urban population is mostly concentrated in a single large city that coexists with other, smaller urban areas.

sorting of the population by skill (human capital) across rural and urban locations. Higher-skilled workers sort into urban areas, where modern production techniques can make use of their abilities, while less-skilled workers remain in the countryside. It appears, however, that Young's approach could be reconciled with the migration equilibrium portrayed in the previous models by positing a continuous skill distribution, where a critical skill level separates migrants and nonmigrants. The worker with this critical skill level would be indifferent to migrating or not migrating, with the higher urban cost of living exactly balancing the higher return to his skill in the city. Workers with skills higher (lower) than the critical level would strictly prefer the urban (rural) area, and the result would be a gap between average living standards in the two areas, as found by [Young \(2013\)](#).

21.4. TENURE INSECURITY: A HALLMARK OF HOUSING MARKETS IN DEVELOPING COUNTRIES

In the analysis in [Section 21.3](#), the cost of urban housing plays a key role in equilibrating rural–urban migration. While the dwellings in the urban housing market were depicted for simplicity as consisting only of land, housing in developing countries is in reality a more complex commodity composed of a bundle of attributes characterizing the structure and the land it occupies, just as in developed countries. In generating prices, housing markets in developing countries function in some respects just like those in developed countries, with the rents or selling prices that dwellings command reflecting the desirability of the attribute bundles they offer. Recognizing this commonality, many researchers have estimated hedonic price models for developing-country housing markets or used related approaches for the purpose of measuring willingness to pay for housing attributes. The results mirror the findings of hundreds of similar studies from developed countries. Studies of this type include those of [Follain et al. \(1982\)](#), [Quigley \(1982\)](#), [Lim et al. \(1984\)](#), [Follain and Jimenez \(1985\)](#), [Daniere \(1994\)](#), [Gross \(1988\)](#), [Lall et al. \(2008\)](#), [Takeuchi et al. \(2008\)](#), and [Brueckner \(2013a\)](#).

Despite some commonality with housing markets in the developed world, markets in developing countries are distinguished in part by the widespread presence of urban slums, which consist of low-quality, often self-constructed, housing much worse in quality than that found in the developed world. For a graphic description of slum conditions in developing countries, see [Marx et al. \(2013\)](#).²⁰ Case studies also suggest that a substantial share of slum residents consists of squatters, who occupy the land without paying compensation to its owner. For example, in the city of Dhaka, Bangladesh,

²⁰ The low quality of the housing consumed by migrants is, of course, not recognized in the models presented in [Section 21.3](#).

squatter settlements are estimated to provide as much as 15% of the housing stock (World Bank, 2007), and the share is probably higher in some other regions.²¹

Squatting represents the most extreme case of another feature that differentiates housing markets in developing countries from those in the developed world: the widespread lack of tenure security, or “land rights” security. Tenure security in a legal sense is completely absent for squatters, although they enjoy some degree of security in practice if the threat of eviction is low. However, owing to underdeveloped legal and land registration systems, tenure is also insecure for many developing-country residents who pay for the housing they occupy but face a “continuum” of property rights, with legal gray areas. This insecurity reduces the incentives to invest in housing improvements by owner-occupiers, an effect that is present in the starkest fashion for illegal squatters, who often live in flimsy shacks constructed from abandoned materials, creating slum conditions. By reducing their attractiveness, tenure insecurity also impedes the market turnover of dwellings, although squatter housing does trade in markets internal to the settlements. Lacking legal status, however, squatters cannot sell their plots to formal users, who would require legal title to proceed with redevelopment of the land. Tenure insecurity also stunts the development of mortgage markets, which require clear title in order to treat a dwelling as loan collateral.²² These obstacles, of course, can be overcome through government programs that transfer land titles to illegal occupants, programs that exist in some cities in the developing world (examples come from Peru and Argentina, as discussed further below).

The remainder of this section explores the issue of tenure insecurity in considerable detail. The next subsection discusses two case studies, for Mali and Vietnam, that show how a property-rights continuum, as mentioned above, works in practice. The discussion then turns to an extensive treatment of the economics of squatting, a polar case of tenure insecurity that has become the focus of a substantial literature.

21.4.1 A continuum of property rights: Mali and Vietnam

Selod and Tobin (2013) provide an extensive discussion of property rights and tenure insecurity in the West African country of Mali, while also developing a theoretical model to explain the observed patterns of security. “Customary tenure,” which is prevalent in rural and periurban areas, is enforced by village chiefs and their councils of elders, following a request for land and a small symbolic payment. Although customary tenure involves no legal documents, it is legally recognized. An “attribution letter” is a document issued at the beginning of a land allocation procedure (which transfers public land to

²¹ With crowding presumably higher in squatter housing, Dhaka’s 15% squatter housing share would translate into a larger share of the population.

²² An upside of tenure insecurity is that easy removal of existing occupants lowers the barrier to land redevelopment when it becomes desirable.

private individuals), and it provides a basic level of tenure security. More formal and secure tenure rights are provided by rural or urban “residency permits,” which provide temporary use rights, and at the end of the continuum is “definitive title,” which provides full property rights. Selod and Tobin state that only 8.5% of owners in Mali hold this title, although the share is higher in cities. They also explain that these different tenure rights are “obtained through complex processes” that are “very costly” for households.

Recognizing these costs, the study authors construct a theoretical model where households invest in tenure security, incurring a cost $C(\pi, e)$ to achieve a probability of π of keeping their property (their housing outlay is lost otherwise). The parameter e measures the household’s idiosyncratic ability to interact with the land administration in pursuit of tenure security, with a higher e reducing C . The model is developed in a spatial context with commuting to a central business district, and it predicts that households are distributed across space according to their values of e , with higher e households (who achieve larger values of π) located closer to the center. The city thus exhibits different tenure-security zones, with the highest security near the center and the lowest security found on the rural fringe, a pattern that roughly matches the one seen in Selod and Tobin’s survey data.

In Vietnam, as explained by Kim (2004), all the land was originally owned by the state, but is being transferred to households through a cumbersome and costly titling process. A title is known as a building occupancy and land-use certificate (BOLUC). To receive a BOLUC, a household must present a collection of “legal papers,” which could include a construction permit, a notarized transfer contract, an occupancy-rights document provided by a local district committee, an inheritance document, or a number of other items. Kim points out that “legal papers” provide some tenure security in the absence of a BOLUC, with such papers appearing to serve a purpose like that of the Malian attribution letter described by Selod and Tobin (2013). In Kim’s sample of properties listed for sale, about one-quarter have a BOLUC, one-quarter have legal papers, and one-third have no evidence of property rights.

To gauge the contribution of property rights to selling prices, Kim estimates an hedonic price regression that includes dwelling characteristics and property-rights dummy variables. She finds, as expected, that possession of property rights of any kind raises the dwelling price (being worth about as much as a telephone connection), with a BOLUC worth more than twice as much as legal papers. Interestingly, possession of both legal papers and a BOLUC is worth more than a BOLUC alone, apparently indicating that even a formal title does not provide indisputable property rights, with the supplementary information given by legal papers adding more evidence. Overall, Kim’s results supplement the anecdotal evidence in Selod and Tobin (2013) by showing that developing countries often have a continuum of property rights, with rights in the lower and middle parts of the continuum still providing benefits.

It should be noted that property rights are limited in a different fashion in countries such as China, where all land is government owned and private developers sign long-term, transferable leases for its use. Although Chinese leases are renewable, users in principle face the possibility of losing their usage rights upon termination of a lease.

21.4.2 The economics of squatting: Theory

The discussion turns now to the case of squatting, where legal property rights are completely absent. Along with other social scientists, economists have expended considerable effort studying the squatting phenomenon, leading to a distinct literature. This literature contains theoretical contributions, which attempt to analyze the incentives and decision-making of squatters and landowners, along with empirical contributions, which rely on scarce data to explore various aspects of the squatting phenomenon. This subsection surveys the theoretical side of this literature, and the next subsection appraises the empirical side.

The existing theoretical models of squatting behavior are all built around the possibility of eviction of squatter households. While eviction is an outcome sometimes faced by legal occupants of rental housing in developed countries (though it is usually circumscribed by renter-protection laws), illegal occupancy in squatter areas makes eviction a more serious threat. In some countries, however, this threat appears to fade rapidly with the length of tenure. In a private conversation, for example, Fernando Cavillieri, Director of the Pereira Passos Institute of the city government of Rio de Janeiro, asserted that squatters in that city must be evicted in their first week of occupancy if they are to be evicted at all. A quick eviction of this type figured prominently in recent news stories describing the eviction of squatters (and an ensuing riot) in a new Rio de Janeiro settlement (Kiernan, 2014). In another private conversation, Zama Mgwatyu of the Development Action Group, a South African NGO, asserted that successful evictions in that country must occur within the first 48 h of occupancy.

While worldwide data on squatter evictions do not exist, these observations are consistent with a view that, despite the threat, actual evictions are relatively infrequent. This view matches the Ecuadorian survey data of Lanjouw and Levy (2002) (discussed in detail below), which show that squatters in Ecuador do not perceive eviction as being very likely. Nevertheless, the eviction threat is the foundation of most research on squatting, although one line of theoretical work (Brueckner and Selod, 2009; Brueckner, 2013b; Shah, 2014) addresses this apparent conflict by explaining how eviction can be both threatened and absent. The argument is that the organizers of squatter settlements ensure that they never expand to a size that would prompt eviction, while also taking other steps to guard against this outcome. By contrast, in the models of Jimenez (1985), Hoy and Jimenez (1991), and Turnbull (2008), eviction is an event that occurs with some

probability, which may depend on how much housing squatters construct on their plots. A large housing investment can raise eviction costs, making eviction less likely. To understand the different perspectives embodied in this work, it is useful to sketch the main features of the individual models.²³

21.4.2.1 The Jimenez model

In Jimenez (1985), a consumer (the head of a household) must decide whether to be a squatter or to rent housing in the formal sector, so squatting is a tenure choice. With formal occupancy, the consumer pays a rental price per unit equal to p_f , consuming h_f worth of housing and x_f worth of the numeraire nonhousing good. Conditional on being a formal resident, the consumer then faces the budget constraint $x_f + p_f h_f = \gamma$, where γ is income. The consumer chooses the two consumption levels to maximize utility $u(x_f, h_f)$ subject to this constraint.

As a squatter, the consumer faces the possibility of being evicted and forced to enter the formal market to secure housing. In this event, the consumer loses whatever funds were spent on squatter housing (at price $p_s < p_f$ per unit) and must make a new housing expenditure. If eviction does not occur, however, the need for this double expenditure does not arise. Note that the Jimenez model, by assuming payment of rent by squatter households, portrays a developed squatter community in which a market for housing exists despite the illegality of the settlement. Even though the land was seized initially without any payment to its owners, trade in squatter dwelling arises subsequently as the community becomes established.

With eviction, the squatter's budget constraint is $x_{fe} + p_f h_{fe} + p_s h_s = \gamma$, where h_s is the amount of squatter housing and where the subscript fe denotes consumption levels in the formal sector following eviction. The second term is the formal housing expenditure necessitated by eviction, whereas the third term is the lost expenditure on squatter housing. Since h_s is never actually consumed when the squatter is evicted, utility in the eviction case is $u(x_{fe}, h_{fe})$. By contrast, if eviction does not occur, then the budget constraint is $x_s + p_s h_s = \gamma$ and utility is $u(x_s, h_s)$. Note that housing consumption equals the squatter level h_s in this case, not h_{fe} .

Eviction occurs with a probability π , which is ultimately endogenous. Expected utility for the squatter household then equals $\pi u(x_{fe}, h_{fe}) + (1 - \pi)u(x_s, h_s)$, the weighted sum of the utilities in the eviction and no-eviction cases, with the weights being the relevant probabilities. Eliminating the x 's using the two budget constraints from above, expected utility can be rewritten as

²³ Eviction can be viewed as an uncompensated "taking" of land, where the absence of compensation is justified by illegal occupancy. However, government-sanctioned takings that occur without (adequate) compensation are common around the world, happening in both developed countries and underdeveloped countries outside squatter settlements.

$$\pi u(\gamma - p_f h_{fe} - p_s h_s, h_{fe}) + (1 - \pi) u(\gamma - p_s h_s, h_s). \quad (21.15)$$

The household chooses h_s and h_{fe} to maximize (21.15). A key feature of this optimization problem is that the squatter housing level h_s is chosen recognizing that it may never be consumed. Given the possibility of this lost expenditure, the consumer will set h_s at a level lower than would be chosen if eviction were impossible. In addition, if the squatter is forced to enter the formal market, the income reduction from the lost squatter-housing expenditure will lead to an h_{fe} smaller than the h_f that would be chosen by a nonsquatter.

Although the formal housing price p_f is exogenously fixed, the squatter price p_s is determined along with the eviction probability π by the interaction of supply and demand forces. On the demand side, consumer “willingness to pay” for squatter housing helps determine its price. Willingness to pay can be found by deriving the price p_s at which a consumer would be indifferent between squatting and formal residence. To do so, note that formal utility is fixed, with its value being determined by γ and p_f , both of which are exogenous. Expected utility as a squatter will also depend on these exogenous variables, but the crucial dependencies are on p_s and π , and an increase in either variable reduces expected utility. With the effects of both variables being negative, keeping expected utility constant at the fixed formal utility level means that p_s and π must vary inversely, with a lower p_s accompanying a higher π , and vice versa. As a result, willingness to pay for squatter housing is a decreasing function of the eviction probability, so that

$$p_s = g(\pi), \quad (21.16)$$

where $g' < 0$. It is easy to see that p_s from (21.16) equals p_f when the eviction probability is zero (which makes formal residence and squatting equivalent), while $p_s < p_f$ holds when $\pi > 0$. In this case, the consumer requires a price discount to bear the risk of eviction.

Two elements interact on the supply side: a limited squatter land area, and fixed amount of government funds available for eviction. The limited squatter land area means that the price of squatter housing increases with the size N of the squatter population, as more squatters compete for the available land. The resulting price equation is written as $p_s = h(N)$, where $h' > 0$. The fixed amount G of government eviction funds means that only G/E squatters can be evicted, where E is the eviction cost per squatter. But with a squatter population of N , the eviction probability is then $\pi = (G/E)/N$. Since a larger N thus reduces π while at the same time raising p_s owing to land scarcity, it follows that p_s and π are inversely related. In other words,

$$p_s = m(\pi), \quad (21.17)$$

where $m' < 0$.²⁴ Therefore, from the supply side, the squatter housing price is a decreasing function of the eviction probability, just as is true on the demand side.

²⁴ Formally, (21.3) follows from rewriting the π equation as $N = (1/\pi)(G/E)$ and substituting this into the price equation $p_s = h(N)$, which yields $p_s = f[(1/\pi)(G/E)] \equiv \Omega(\pi)$. Since $f' > 0$, it follows that $\Omega' < 0$.

The demand relationship in (21.16) and the supply relationship in (21.17) jointly determine the equilibrium values of the squatter housing price p_s and the eviction probability π . The equilibrium corresponds to the intersection of the two curves given by (21.16) and (21.17). Once the equilibrium π has been determined from this intersection, the equilibrium size for the squatter population is found via the formula $\pi = (G/E)/N$. Since both demand and supply relationships are downward sloping, they may have several intersections, which means that multiple equilibria may exist. Some of these equilibria will be unstable.

The building blocks of the Jimenez model are thus consumer indifference between squatting and formal residence on the demand side, and limited land for squatting along with fixed eviction funds on the supply side. The price p_s , the eviction probability π , and the squatter population size N all adjust to make consumers indifferent between the tenure modes, to ensure that all eviction money is spent, and to allow squatters to fit in the available land area. Comparative-static analysis of the model shows that an increase in eviction spending G reduces N and p_s while raising π , all natural conclusions. The effects of an increase in income y are ambiguous.

21.4.2.2 The Hoy-Jimenez and Turnbull Models

Rather than having the government evict squatters using a budget of fixed size, Hoy and Jimenez (1991) and Turnbull (2008) assume that evictions are instead carried out by landowners in response to emerging development opportunities for particular plots of land. These opportunities arise randomly, making eviction on a particular plot a stochastic event.

In the Turnbull model, the posteviction fate of squatters is harsher than in the model of Jimenez (1985), with housing consumption reduced to zero following eviction. Expected utility is then $\pi u(x_s, 0) + (1 - \pi)u(x_s, h_s)$, and the squatter budget constraint is $x_s + p_s h_s = \gamma_s$, where γ_s is squatter income. Since squatter housing expenditure is lost with eviction, h_s is smaller when the eviction probability is higher. This relationship is written $h_s = t(\pi)$, where $t' < 0$. The squatter housing price, which plays no important role in the model, can be viewed as fixed. In addition, the tenure choice aspect of the Jimenez model is absent.

Turnbull assumes that the net revenue r that can be earned by developing a plot is a continuous random variable, which is independent across plots. Landowners each own just a single plot, and any particular landowner will evict squatters when the r realization for the plot exceeds the cost of eviction. This cost depends positively on the amount of housing on the plot, which must be cleared following eviction, being written as $c(h_s)$, with $c' > 0$. The developer thus evicts the squatter when the random development revenue r exceeds $c(h_s)$ and does not evict the squatter otherwise. Eviction is then less likely when $c(h_s)$ is large and hence when h_s is large. As a result, the eviction probability can be written as a decreasing function of h_s , with $\pi = q(h_s)$, where $q' < 0$.

Since h_s is a function of π , while π depends on h_s , the housing consumption level and eviction probability are jointly determined. The mutually consistent equilibrium values

of h_s and π must satisfy both of these relationships, with $h_s = t(\pi)$ and $\pi = q(h_s)$ holding. The values thus lie at the intersection of the two curves defined by these equations. As in the model of Jimenez (1985), the curves are both downward sloping, raising the possibility of multiple equilibria, some of which are unstable. Turnbull's comparative-static analysis shows that h_s falls and π rises when favorable development returns become more likely or when squatter income falls.

Whereas landowners are atomistic in Turnbull's model, with each owning a single plot, ownership is concentrated in the Hoy–Jimenez setup, with one landowner owning all the land occupied by squatters. In addition, the random development opportunities are binary in nature. For each plot, a development opportunity yielding a fixed net revenue R arises with probability θ , and no opportunity arises with probability $1 - \theta$.

As in the Turnbull model, the eviction cost is equal to $c(h_s)$, with $c' > 0$. In addition, housing consumption again depends on the eviction probability π via the function $h_s = t(\pi)$, with $t' < 0$. But in contrast to the Turnbull model, where π for any given plot reflects the randomness of development revenue for that plot (and hence the randomness of the development decision), π in the Hoy–Jimenez model is determined by the landowner's choice of *the fraction of plots on which eviction will occur*. From the squatter's perspective, eviction looks random because no squatter knows in advance whether his/her plot will be targeted for eviction. The eviction share, however, is chosen in advance by the landowner, recognizing that the identities of the particular plots on which eviction occurs will depend on the random arrival of development opportunities.

The main question of interest for Hoy and Jimenez is the relationship between the landowner's chosen π and the probability θ of a development opportunity, quantities that they show need not be equal. First, it is clear that no plot with a favorable opportunity for development will be passed over for eviction, while eviction occurs on some plot that lacks such an opportunity. Costs would be unaffected by switching the location of eviction, while revenue would rise. But it may be optimal for the landowner to evict squatters on all plots with development opportunities *while also evicting squatters on plots without them*, so that $\pi > \theta$. The reason is that the larger π will retard squatter investment (h_s) on all plots, limiting expected eviction costs. In this case, Hoy and Jimenez say that "apparently superfluous" evictions occur. A third possibility is that π is set low enough so that evictions do not occur on some plots with development opportunities, so that $\pi < \theta$.

Hoy and Jimenez's main conclusion is that either of these possibilities could indeed be optimal, with π either smaller or larger than θ , or possibly equal to it.²⁵ Comparative-

²⁵ Since eviction cost depends on h_s and h_s depends on π , the cost can be written as $k(\pi)$, a decreasing function, so that the expected cost per plot equals $\pi k(\pi)$. The expected profit per plot is then $\pi R - \pi k(\pi)$ when $\pi < \theta$ and $\theta R - \pi k(\pi)$ when $\pi \geq \theta$. Hoy and Jimenez show that the maximizing value of π can lie in either of these ranges.

static analysis shows that when superfluous evictions occur, marginal increases in R or θ have no effect on the chosen π . When the optimal π is less than θ , however, a marginally higher R raises π , although an increase in θ still has no effect.

Unlike in the Hoy–Jimenez setup, evictions in the Turnbull model are made atomistically, plot by plot, which means that superfluous evictions (where no development revenue is earned) will never occur. In the Hoy–Jimenez model, by contrast, the single landowner may want to carry out evictions on plots without development opportunities in order to raise the general threat of eviction, thus reducing housing investment (and making eviction easier) throughout the squatter area.

21.4.2.3 The Brueckner–Selod model

The model of squatting of Brueckner and Selod (2009) is centered around the eviction issue, but the approach is entirely different from the approaches of the previous authors. As in the Hoy–Jimenez and Turnbull models, development of the land occupied by squatters is desirable when the revenue earned is high. But instead of following previous authors by saying little about the market for posteviction developed plots, Brueckner and Selod assume that such plots *are rented out in the city's formal housing market*. For simplicity, housing capital is absent, with land consumption representing housing. So when eviction removes squatters from the land, the vacant plots are rented and occupied by formal residents. Since squatters and formal residents divide a fixed total land area, the presence of squatter settlements “squeezes” the formal market, raising the formal rent per unit of land, denoted p_f .

Eviction is desirable when p_f exceeds the eviction cost per unit of land. While this “eviction condition” follows the spirit of the other models, the key difference in the Brueckner–Selod model is the existence of a squatter organizer, who ensures that the eviction condition is not met and hence that eviction never occurs. The organizer achieves this goal by limiting the squeezing of the formal market, preventing escalation of p_f , and by controlling eviction costs. As explained above, this structure helps to explain real-world patterns, in which the threat of eviction exists but its occurrence is relatively infrequent.

The eviction costs that the organizer attempts to control depend in part on the “defensive expenditures” undertaken by squatters, which raise the cost of eviction. These expenditures, which are dictated by the squatter organizer, could go partly toward political lobbying intended to build support for the squatter community. The expenditures could also support a squatter security force for defending the settlement, or they could represent foregone labor income as squatters spend time at home to defend their plots rather than working at full capacity (Field (2007) provides evidence on such behavior, as discussed below). Jimenez (1985) briefly includes defensive expenditures in his model, but he argues that they represent a public good exploitable through free riding, implying

that the equilibrium level of such expenditures will be zero. With the squatter organizer dictating individual defensive expenditures, a positive level is sustainable.

The eviction cost thus depends on defensive expenditures per household, which are denoted by A . In addition, the size of the squatter population, N_s , affects eviction costs in a positive direction. With a larger total squatter population, the political outcry caused by eviction is more substantial, making eviction more costly. These relationships are captured by the eviction-cost function $e(A, N_s)$, which gives the eviction cost *per unit of land*. The $e(\cdot)$ function is increasing in both its arguments. In order for eviction not to be worthwhile for landowners, the posteviction return to the land, given by the formal price p_f , cannot be larger than the eviction cost per acre. Formally, this “no-eviction” constraint is written

$$p_f \leq e(A, N_s). \quad (21.18)$$

The squeezing process determines p_f in (21.18), as follows. Individual land consumption for squatter households is denoted by h_s , which implies that the total land area occupied by squatters equals $N_s h_s$. With the overall land area of the city fixed at $\bar{L}$, the remaining formal land area equals $\bar{L} - N_s h_s$. The fixed formal population must fit in this area, which requires the formal price p_f to adjust so as to equate the total demand for land by formal residents to the available area. This total demand depends on the size of the formal population, which is fixed at $\bar{N}_f$, and on the individual demand for land, which is given by the downward-sloping demand function $h_f = d_f(p_f)$. Total demand then equals $\bar{N}_f d_f(p_f)$, so that the condition

$$\bar{N}_f d_f(p_f) = \bar{L} - N_s h_s \quad (21.19)$$

ensures that the formal residents fit into the available land area.

The last elements of the model are the squatter utility function, $u(x_s, h_s)$, and the budget constraint. In contrast to the other models, squatters are assumed to incur no direct cost for the land they occupy, which is invaded and occupied with no payment to anyone. Squatters do, however, pay for defensive expenditures, as dictated by the squatter organizer. As a result, their budget constraint is $A + x_s = \gamma_s$, so utility can be written

$$u(\gamma_s - A, h_s). \quad (21.20)$$

While the squatter organizer dictates the level of A , he/she also controls the sizes of squatter plots, dictating the individual land consumption levels h_s . In addition, he/she has control over the size of the squatter population, N_s , having the power to limit the number of households participating in the land invasion. The organizer thus controls A , h_s , and N_s , and he/she chooses the levels of these variables to maximize individual squatter utility, as given in (21.20). The constraints for the optimization problem are the formal market-clearing condition (21.19) and the no-eviction constraint (21.18). It is easy to

see that this constraint will bind at the solution, so that landowners are indifferent to evicting or not evicting the squatters.²⁶

Given the complexity of the model, general comparative-static analysis of the squatter equilibrium is not feasible. However, using common functional forms,²⁷ the equilibrium solution can be computed, showing how the decision variables respond to changes in the exogenous variables. One surprising feature of the solution (a consequence of the assumed functional forms) is that squatters occupy exactly half of the city's land area regardless of the values of the other parameters. With land supply to the formal sector thus effectively fixed, the formal price depends only on the strength of formal demand and not on squatter characteristics such as income y_s . Squatter income does, however, affect N_s and h_s , which fall and rise, respectively, with y_s in an offsetting fashion so as to keep the total squatter land area constant.

If all squatters were simultaneously switched to formal residency, being required to pay for their land, they would individually be worse off and the formal residents would be better off. The formal residents gain because the squatter group squeezes them less when formalized than it did originally, allowing formal land consumption to rise. Although formalized squatters are worse off, the analysis shows that formal residents could compensate them for their losses while still coming out ahead. This potential Pareto improvement shows that the original squatting equilibrium was inefficient.

Brueckner (2013b) extends this model by assuming that the city has multiple squatter organizers who are rent-seekers rather than benevolent agents. In addition to collecting defensive expenditures, the organizers require squatters to pay rent, which they pocket as income. The article characterizes the squatting equilibrium for this case and presents a variety of comparative-static results.

21.4.2.4 The Shah model

Whereas the Brueckner–Selod model portrays the squatters as occupying privately owned land, much squatting in reality occurs on government-owned land. Shah (2014) adapts the Brueckner–Selod approach to deal with this important alternative case. The loss from squatter occupation of government-owned land could involve forsaken benefits from blocked infrastructure projects, such as enlargement of an airport ringed by squatter settlements (Mumbai, India, represents such a case), or forgone revenue

²⁶ In order for the organizer to face a willing supply of squatters, allowing him/her to control N_s , the utility they achieve must be greater than the utility level reached in the rural area that supplies urban migrants. However, under an alternative version of the model, the organizer cannot control N_s , although he/she is still able to dictate A and h_s . In this case, squatters enter the city until the maximized utility is pushed down to the rural level.

²⁷ The assumptions are that both squatters and formal residents have Cobb–Douglas preferences and that the eviction-cost function is multiplicative in A and N_s .

from sale of the land to the private sector for residential, commercial, or industrial development. Shah (2014) represents such losses in a simple, stylized fashion by assuming that vacant government-owned land generates open-space benefits for formal households (being a city park), which are reduced when the land is partly occupied by squatters.

For simplicity, housing (land) consumption for the formal households is exogenously fixed, as is the formal housing price. The disposable income of a formal household is then $\tilde{y}_f = y_f - \bar{p}_f \bar{h}_f$, where the bars denote fixed values. With h_f fixed at $\bar{h}_f$, formal utility can be written as a function of only nonhousing consumption x_f and open space, denoted ℓ . Formal utility is then $u_f(x_f, \ell)$. Letting $\bar{L}_G$ denote the total amount of government-owned land devoted to parks, the amount of open space available after the squatter land invasion is given by $\ell = \bar{L}_G - N_s h_s$. With the budget constraint given by $x_f = y_f - \bar{p}_f \bar{h}_f = \tilde{y}_f$, formal utility then equals $u_f(\tilde{y}_f, \bar{L}_G - N_s h_s)$.

If the squatters are evicted, the full amount of open space $\bar{L}_G$ can be enjoyed, but eviction costs must be incurred. The government pays these costs, but it finances the expenditure with taxes on the formal households. The total cost of evicting the squatters is the cost per unit of land, $e(A, N_s)$, times the amount of land occupied, $N_s h_s$. Therefore, the eviction tax on each formal household is equal to

$$\frac{N_s h_s e(A, N_s)}{\bar{N}_f}. \quad (21.21)$$

If the utility of formal residents when eviction is carried out exceeds their utility with the squatters present but no eviction tax levied, then eviction will occur. Thus, to avoid eviction, the squatter organizer must ensure that the following no-eviction constraint is satisfied:

$$u_f\left(y_f - \frac{N_s h_s e(A, N_s)}{\bar{N}_f}, \bar{L}_G\right) \leq u_f(\tilde{y}_f, \bar{L}_G - N_s h_s). \quad (21.22)$$

The constraint says that formal utility with eviction is less than or equal to formal utility with the squatters in place. As before, the organizer sets A , h_s , and N_s to maximize squatter utility, $u_s(y_s - A, h_s)$, but he/she now takes into account this new form of the eviction constraint. With squeezing of the formal housing market not an issue, a constraint such as (21.19) does not apply.

As in the case of the Brueckner–Selod model, general comparative-static analysis of the Shah model is not feasible. This obstacle is still present with the simplification of Cobb–Douglas preferences, but when the formal and squatter utility functions are linear, results can be derived. An increase in the number of formal households $\bar{N}_f$, which lowers the per capita eviction tax, raises N_s as the organizer enlarges the squatter population to deter now cheaper eviction. This increase allows defensive expenditures A to be reduced,

but it also requires a smaller h_s . An increase in the valuation of open space by formal residents, which again raises the incentive to evict squatters, has the same effects.²⁸

Like Brueckner and Selod, Shah investigates formalization of squatters, who are given title to the land they occupy in return for a rental payment. These payments, which can be transferred to the formal households, are sufficient to compensate them for the loss of open space only when their valuation of such space is sufficiently high (note the contrast to Brueckner and Selod's unconditional formalization result). Finally, Shah investigates a hybrid model where squatting occurs both on government-owned land and on private land, with the squatters squeezing the formal residents as before. She shows that it is optimal for the squatter organizer to equally split his/her population, with half squatting on private land and half squatting on government-owned land.

21.4.3 The economics of squatting: Empirical work

Despite substantial interest in squatting, empirical work on this topic has not been particularly extensive, mainly as a result of limited data availability. One group of articles focuses on the effect of the tenure insecurity faced by squatters on the rents and values of squatter dwellings and on the ease of transferability through market transactions. Jimenez (1984), Friedman et al. (1988), Lanjouw and Levy (2002), and Kapoor and le Blanc (2008) study these effects. A pair of additional articles by Field (2005, 2007) focuses on the effect of tenure insecurity (and the resulting need to defend the squatter's dwelling) on labor force participation as well as the effect on investment in property improvements. An article by Hidalgo et al. (2010) studies land invasions by squatters in Brazil and relates their occurrence and size to a number of variables measuring local economic conditions. The connections between these empirical studies and the theories surveyed in Section 21.4.2 will be noted as the discussion unfolds.

Jimenez (1984) compares the actual rent that a squatter dwelling commands with the predicted rent for a dwelling with the same characteristics in the formal market. He expects to find a squatter discount relative to the formal rent, reflecting tenure insecurity. This prediction follows formally from equation (21.16), which showed that p_s is less than p_f when the eviction probability is positive.

²⁸ The model Shah analyzes is actually slightly different from the one described above. Rather than using an eviction-cost function that gives the cost per unit of land, she relies on a total eviction-cost function, which gives the cost of evicting all the squatters, a function that is written as $E(A, N_s)$. Therefore, her eviction-cost tax, equal to $E(A, N_s)/\bar{N}_f$, does not depend on h_s , in contrast to the formulation in (21.11). The (inaccurate) description in the text is meant to maintain comparability with the discussion of the Brueckner-Selod model, and Shah's comparative-static results would not necessarily emerge under the text formulation. Her particular results also require the assumption that the cross partial derivative $E_{N_s, A}$ is zero or negative. This condition, which indicates that a large squatter population reduces the marginal benefit from defensive expenditures, is natural.

Jimenez combines actual data on squatter rents and dwelling characteristics for a 1983 Philippine sample with an estimated hedonic price function for formal-sector housing. This function relates formal rent to the number of rooms, measures of structure quality (including the presence of toilet facilities), a water availability index, indicators of phone and electricity access, and some neighborhood characteristics. For each squatter dwelling, the predicted formal rent is generated by substituting the dwelling's characteristics into the formal hedonic price function. The ratio of the predicted formal rent and the actual rent for the squatter dwelling is computed and then averaged across the squatter dwellings in the sample. The average value of the ratio is 1.177, indicating an 18% formal premium, or conversely a 15% discount relative to the formal rent for a squatter dwelling.²⁹ These findings mirror the results of Kim (2004) for Vietnam, which showed the value of different degrees of tenure security in the formal market.

Friedman et al. (1988) use the same dataset to answer the same question in a slightly different way. Instead of just estimating a formal hedonic price function, they estimate two hedonic functions, one for the formal sector and one for squatters, while also estimating a pooled hedonic price function, with a dummy variable representing squatter dwellings in the pooled sample. In the case of renters, the separate formal and squatter hedonic functions are not statistically distinguishable, allowing the pooled function to be used instead and the squatter discount to be inferred from the squatter dummy coefficient. This coefficient shows a squatter discount of around 10%, but one that is not statistically significant. The separate hedonic functions are statistically different for the case of owners, so both functions must be used to predict the squatter discount. The results show a 19%, statistically significant discount.³⁰ Thus, as in Jimenez (1984), the squatter discount is larger for owners.

Kapoor and le Blanc (2008) carry out a related exercise that focuses on the difference between the “rate of return” on a dwelling in the formal and squatter sectors. This rate of return is equal to the dwelling's rent-to-value ratio, which is computed using a 2002 survey from Pune, India. With a squatter dwelling under threat of seizure by the landowner, its rent flow is more uncertain than that of a formal dwelling, implying the rate of return for the squatter dwelling should be higher. In other words, the selling price should be a smaller multiple of the current rent than for a formal dwelling. Kapoor and le Blanc's results show average rates of return for formal and squatter housing of 6.7% and 8.1%,

²⁹ The same exercise is carried out for squatter dwellings that are owner occupied rather than rented using a separate hedonic price function estimated for formal owner-occupied housing. In this case, the average ratio is 1.578, which implies a 37% discount for squatter owner-occupied dwellings.

³⁰ The method is to compute the predicted formal selling price of a squatter dwelling using formal hedonic function and then to compute the predicted selling price of the dwelling as a squatter dwelling using the squatter hedonic function. The ratio is then computed and averaged across squatter dwellings, yielding an average value of 1.23, which implies a 19% squatter discount. Note that the only difference relative to Jimenez (1984) is the use of the predicted squatter price rather than the actual price in the computation.

respectively, with the difference being statistically significant. The squatter rate of return is thus higher, as predicted.³¹

Do these empirical results shed any light on the relevance of the theoretical models discussed in [Section 21.4.2](#)? The main lesson of the results is that the uncertain tenure security of squatters indeed matters for rents, values, and rates of return. Since eviction risk is a key element of the models of [Jimenez \(1985\)](#), [Hoy and Jimenez \(1991\)](#), and [Turnbull \(2008\)](#), the perspective taken in these models is thus supported by the results. The results are less favorable for [Brueckner and Selod \(2009\)](#) and [Shah \(2014\)](#), whose prescient squatter organizer eliminates eviction risk, although these models may still have relevance.

[Field \(2007\)](#) explores the effect of tenure insecurity on labor supply, arguing that the squatter's need to "defend" the household plot may require a physical presence that limits labor force participation. Field uses the varying incidence across households of the Peruvian national land titling program (implemented over the 1995–2003 period) to generate tenure–security differences that then yield observable differences in labor supply in her sample. The empirical setup is a difference-in-differences framework where labor supply is regressed on squatter status and on the interaction of squatter status and a dummy variable indicating coverage of the household's neighborhood by the land-titling program, whose spatial scope was systematically expanded over the 1995–2003 period. The results show a 13 h weekly work reduction associated with nontitled squatter status, reflecting a 14% labor supply reduction for the household. By showing the large burden of "defensive expenditures" in the form of substantial forgone labor income, Field's findings lend some credence to the structure of the Brueckner–Selod and Shah models.

[Field \(2005\)](#) applies a similar framework to study the effect of reduced tenure insecurity on residential improvements. Recalling that a higher eviction probability reduces housing investment in the models of [Jimenez \(1985\)](#), [Hoy and Jimenez \(1991\)](#), and [Turnbull \(2008\)](#), Field's inquiry is closely linked to these squatting theories. Her empirical setup is again a difference-in-differences framework that relies on the spatial expansion of the Peruvian land titling program. The results show that when a squatter gains title to a plot, housing improvement spending (mainly in the form of small additions) rises by 68%. [Galiani and Schargodsky \(2010\)](#), again using a natural experiment involving random allocation of titles, find a similar effect of titling on dwelling quality in Argentina, presumably a reflection of higher investment. They also show that titling reduces household size and raises child education levels.³²

³¹ Each survey respondent was asked to state both the rent and the sales value that would be commanded in the market for a dwelling similar to theirs. Both renters and owners were asked the same questions, so the rent and value, and hence the rate of return, are available for dwellings in both groups.

³² See [Galiani and Schargodsky \(2010\)](#) for a broader survey of studies on the effects of property rights. See also [De Soto \(2000\)](#), who argues that extraction of housing equity is a major source of entrepreneurial capital that is denied to residents of developing countries who lack property rights.

The main goal of [Lanjouw and Levy \(2002\)](#) is to investigate the connection between tenure security and the ability to transfer a dwelling via sale or rent to another household. But as a prelude to this inquiry, the study authors provide unusual evidence about the perceived likelihood of eviction and the role of squatter organizers. Using 1995 survey data from Ecuador, Lanjouw and Levy report squatter responses to a question about the likelihood of eviction, with possible answers being that eviction is “sure or very possible,” “possible,” “not very possible,” or “impossible.” Among 142 respondents, none of whom held title to their plot, the first answer is never given, while the “possible” answer is given by only 3.5% of the respondents. This evidence appears consistent with the Brueckner–Selod/Shah view of eviction as a very unlikely event.

When the respondents are divided into groups whose squatter community has an organizer (111 respondents) or does not have an organizer (31 respondents), the total number of responses in the “sure or very possible” or “possible” categories remains below 3.5% for each group. By contrast, the split between the last two categories (“not very possible” or “impossible”) differs across the groups. For respondents in communities with an organizer, 83% say that eviction is “impossible” (with 13.5% saying it is “not very possible”), while in communities without an organizer, 58% say eviction is “impossible” (with 39% saying it is “not very possible”). Therefore, the presence of a squatter organizer substantially increases a squatter’s perception of freedom from eviction, consistent with the view taken in the Brueckner–Selod and Shah models. These results are reaffirmed in a probit regression where other determinants of the perceived impossibility of eviction are included (possession of an ownership document, which has a positive effect; squatting on private land, which has a negative effect).

To gauge the determinants of the transferability of a dwelling, Lanjouw and Levy run a probit regression using responses to a question asking whether the household “can sell or rent” its dwelling in an arm’s length transaction, where the other party is a stranger. The age (and hence established nature) of the community raises transferability, as does the presence of a squatter organizer, with the organizer’s effect being stronger the older the community.³³ Thus, the main implication of Lanjouw and Levy’s results for theoretical perspectives on squatting is the revealed importance of the squatter organizer. The presence of an organizer reduces the perceived risk of eviction, as argued in the Brueckner–Selod and Shah models, and that presence also facilitates the operation of an internal market for squatter housing, presumably by giving more organizational structure to the squatter community.³⁴

³³ [Di Tella et al. \(2007\)](#) explore another effect of tenure security from receipt of a land title: the effect on beliefs about the fairness of the operation of the market system. Receipt of a title strengthens such beliefs.

³⁴ Transferability is also higher when adult males are present in the household and when a title is held, and a number of other covariates have effects.

The final empirical study to be considered, that of [Hidalgo et al. \(2010\)](#), was mentioned above in the discussion of rural–urban migration. It investigates the occurrence and extent of squatter land invasions using community-level data. This article’s remarkable dataset tabulates the occurrence of distinct land invasions in all Brazilian cities over the 1988–2004 period. Almost 5300 invasions are observed, and when an invasion occurs, the dataset indicates the number of participating households. The main hypothesis tested by the authors is that the occurrence of a land invasion is more likely (and its size greater) the lower is the agricultural income in the countryside surrounding a city, as measured by crop yields per hectare. A concern about endogeneity of agricultural incomes is the motivation for the use of an instrumental variables approach, with a rainfall measure serving as the instrument. Hidalgo et al. argue that reverse causality may arise because land invasions divert people from agricultural employment, possibly affecting the harvest and reducing crop yields, thus lowering agricultural incomes.

Estimates from a linear probability model show that, as expected, the occurrence of a land invasion becomes more likely as agricultural income falls. A different specification using a variable equal to the count of distinct land invasions in a city also shows the same negative effect of agricultural income, as does a regression where the dependent variable is the number of invading households.³⁵

These findings make intuitive sense, and they also match theoretical results from [Brueckner and Selod \(2009\)](#) and [Shah \(2014\)](#). In the free-migration versions of those models, where the size of the squatter population cannot be controlled by the organizer and expands to equalize rural and urban utilities, a lower rural utility leads to a larger equilibrium squatter population.

21.4.4 Avoiding the formation of squatter settlements or relocating them

The squatting literature contains little discussion of government policies that are designed to limit squatter populations. One policy would consist in withholding public services such as water and electricity from squatter areas, which would reduce their attractiveness and limit inward migration. [Feler and Henderson \(2011\)](#) study this question by exploring how the availability of water connections in the informal housing areas of Brazilian cities affects population growth at the city level. They find that a higher share of informal dwellings with water connections leads to faster urban population growth, as intuition would predict. Feler and Henderson also investigate strategic interaction among cities in the choice of their water-connection shares, finding evidence that cities look to their neighbors in deciding what fraction of informal dwellings to service. Note that the

³⁵ The study authors also present further specifications including interaction terms where the effect of agricultural income depends on the inequality of land holdings in the region. The results show that a decrease in agricultural income raises the likelihood and the size of land invasions more when land holdings are more unequal.

findings of this article relate to the role of public services in attracting migrants, as discussed in [Section 21.2](#). More broadly, the article suggests that policies such as the withholding of utility connections that are designed to limit urbanization may encourage slum conditions in areas where rural–urban migrants, undeterred by the policies, nevertheless arrive. In other words, poor living conditions are fostered in underserved areas that still attract migrants.

While policies designed to make squatting less attractive can limit rural–urban migration, relocation of slums and squatter areas once they come into being, which allows land to be redeveloped to a higher use, is another policy pursued in some countries. [Lall et al. \(2008\)](#) and [Takeuchi et al. \(2008\)](#) use Indian data to study the preconditions for successful policies of this type. They estimate slum dwellers' valuations of housing and neighborhood characteristics, with the goal of gauging what characteristics new (titled) settlements must have in order to make relocation welfare improving.³⁶

21.5. PROVISION OF AFFORDABLE HOUSING IN DEVELOPING COUNTRIES

The squatting phenomenon, extensively discussed in the preceding section, is a symptom of the lack of affordable housing in the cities of developing countries. In other words, the initial households arriving in a squatter settlement, who occupy vacant land, are solving an affordability problem by appropriating the land for free. Policies that increase the availability of affordable housing can therefore reduce the extent of squatting and slums in developing countries while also benefiting residents in the formal housing sector, who may then gain access to better shelter at a lower cost. This section focuses on three inter-related issues related to the provision of affordable housing in developing countries: the extent of overall investment in housing, land-management policies that regulate transactions and land use, and investment in complementary infrastructure and services.

21.5.1 Investment in housing

21.5.1.1 *The responsiveness of supply*

Provision of affordable housing requires that housing supply responds to the increase in demand caused by rural–urban migration and rising incomes. In a classic study of supply responses, [Burns and Grebler \(1976\)](#) examine the share of housing investment (measured by new residential construction) in GDP, employing data from 39 countries. By regressing the investment share on GDP per capita, the change in population, and a measure of urbanization, they find that the share of housing investment in GDP increases at an early stage of development but declines past about \$6500 per capita GDP (2005 US dollars).

³⁶ For other approaches to fostering land development and redevelopment in developing countries, see [Lozano-Gracia et al. \(2013\)](#).

Subsequent to this study, there has been surprisingly little empirical work on assessing the supply elasticity of housing in response to demands from urbanization.

A new study by Dasgupta et al. (2014) provides more evidence on housing supply by building a new dataset on housing investment covering over 90 countries worldwide. The dataset explicitly accounts for investments by households, the government, and the private sector. This database uses the variable “ownership of dwelling, value added” in the National Account Statistics, and covers buildings that are used entirely or primarily as residences (including houseboats, barges, mobile homes, and caravans), together with any associated structures, such as garages, and all permanent fixtures customarily installed in residences. An important contribution of the study is the inclusion of informal housing, a necessity for accurate measurement given that formal housing constitutes a very small share of the housing stock in many developing countries. For example, in many African cities and towns, less than 10% of the population lives in formal housing.

The evidence from the study shows that investments in housing pick up as countries approach middle incomes. In fact, housing investment follows an S-shaped trajectory, with low-income economies having small income elasticities as a result of both supply and demand constraints. On the supply side, rigidities in the supply of materials, organization of the construction industry, and nascent markets for land transactions constrain housing expansion in urban areas. On the demand side, when incomes are low at early stages of development, the claims of other types of consumption expenditure, such as food expenditure, tend to dominate. But the empirical evidence shows that housing investment takes off at income levels of about \$3000 per capita (2005 US dollars) and then slows at income levels around \$36,000 per capita (2005 US dollars).

The study shows that, over time, the share of housing investment in GDP has been steadily increasing for developing and middle-income countries. Between 2001 and 2011, housing investment averaged 4.56% of GDP in low-income countries (up from 2.51% between 1960 and 1971), 6.06% in middle-income countries (up from 4.38%), and 9.12% in upper-middle-income countries (up from 5.11%). In contrast, housing investment as a share of GDP has been gradually declining in high-income (OECD) countries, having peaked in the 1960s (see Table 21.1).

The study’s second main finding is that the elasticity of housing investment with respect to a country’s urbanized population has steadily increased for developing countries and is estimated to be around 4 in 2011. This pattern is seen in the first panel in

Table 21.1 Decadal average of housing supply (percentage of GDP) across country income categories.

Income category	1961–1970	1971–1980	1981–1990	1991–2000	2001–2011
Low income		3.00	3.58	4.79	4.56
Lower middle income	4.38	4.97	5.41	6.10	6.06
Upper middle income	5.11	8.49	9.14	8.86	9.12
High income: OECD	11.29	9.87	8.06	6.03	5.71

Source: Dasgupta et al. (2014); countries classified as per World Bank classification.

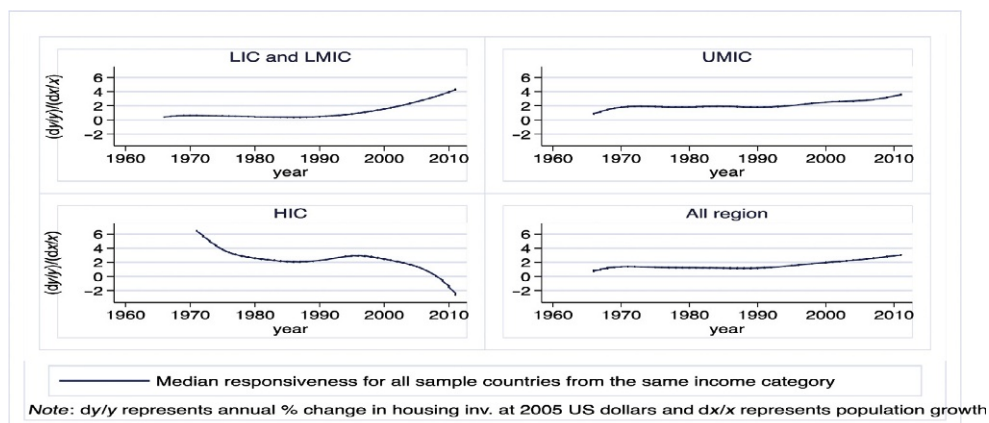


Figure 21.2 Housing supply responsiveness to urban growth across income groups.

Figure 21.2, which plots median elasticity values for low-income and lower-middle-income countries over the last 40 years. Upper-middle-income countries also experienced a rising elasticity, although the trend is less pronounced than in low-income countries (see Figure 21.2). By contrast, the elasticity has been declining for high-income (OECD) economies, where it peaked in the 1970s. This pattern reflects a moderate urban rate of growth of 2.31 during the 1961–1970 period along with a large 11.29% investment share, combined with a falling investment share (see Table 21.1) and moderating urban growth after 1970. Overall, these numbers are encouraging because they show a rising housing-supply response to urbanization in developing countries, demonstrating that market forces are reacting to surging housing demand caused by rural–urban migration and rising incomes.

To better understand the timing of housing investment in response to urbanization, Dasgupta et al. (2014) also develop the following typology of investment patterns: (a) leading, where housing investments occur ahead of urbanization; (b) synchronous, where investment occurs contemporaneously with urbanization; and (c) lagging, where housing investment trails urbanization. They find that leading countries tend to have higher magnitudes of foreign and domestic investment, as well as larger current-account trade deficits, than lagging countries. Not surprisingly, OECD countries form the majority in the leading group, while most parts of the world appear to have a synchronous relationship between investments and urbanization. However, housing investment appears to follow demand with a lag of 8 or 9 years in sub-Saharan Africa, indicating an inadequate supply response to urbanization in this part of the world.

21.5.1.2 Some determinants of housing affordability

Beyond the responsiveness of supply, affordability depends on household incomes and the cost of housing, which in turn depends on construction costs, including the price of cement and labor costs, on the market structure of the construction industry, and

on building standards and other regulations. In addition, affordability depends on the trade-offs households are willing to make between housing and other necessities. Unfortunately, there is very little empirical evidence regarding housing affordability or its determinants. But [Collier and Venables \(2013\)](#) argue that affordability of construction is a requirement for widespread investment in housing. They provide the following illustrative example:

In Dar es Salaam, the typical rental rate per room is around \$10 a month, and in Dakar around \$16. So, a modest four-room home (equivalent to the “two-up, two-down” of 19th-century Britain) would be affordable, albeit perhaps for multioccupant use, in the repayment range of \$500–\$800 a year. What this implies for a viable purchase cost depends on the real interest rate and terms of financing, but it is difficult to see such a repayment rate supporting a home costing more than around \$15,000. This, of course, includes the price of land, which in Dar es Salaam on the informal market is currently around \$5,000 for a plot of 300 square meters. Clearly, such a plot could support more than one small house, but given current land availability, land costs per house could not be reduced much below \$2,000.

Affordability would thus require construction costs of no more than \$13,000, and although this level might not be attainable under normal circumstances, Collier and Venables point to “a pilot project by Tanga Cement in Dar es Salaam, using modern techniques of pre-casting and four-storey construction” that achieved it.

In addition to income considerations, affordability must be assessed relative to the share of budgets that ordinary households are willing to devote to housing. One of the few systematic assessments of spending on housing is reported in a recent World Bank study on sub-Saharan Africa ([Lozano and Young, 2014](#)). The study finds that, across the region and income classes, household expenditures on housing were low, averaging around 12% of the budget (see [Figure 21.3](#)). This low share is due to very high levels of spending on food, which reach 60% for the poorest quintile, reflecting an early stage

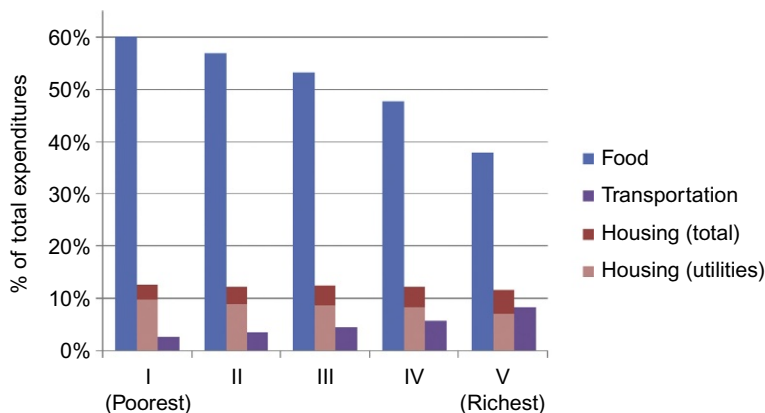


Figure 21.3 Urban household expenditures across sub-Saharan Africa.

of economic development. But even the richest quintile in sub-Saharan Africa spends a larger percentage on food than the average household in South Korea spent in 1975 (when it was a low-income country). Given the severe budget constraints reflected in these numbers, it is difficult to expect a significant increase in housing investment in Africa without commensurate rapid growth in incomes. This lesson extends beyond the African context, with other cross-country evidence showing that households in low-income countries spend around 47% of their total budgets on food (Regmi et al., 2001). Food spending consumes a smaller 29% of household budgets in middle-income countries and just 13% in high-income countries. Correspondingly, income elasticities for food tend to shrink as country incomes grow.

21.5.2 Land-management policies

Policies regulating land transactions and land use constitute another factor affecting provision of affordable housing. Such regulations are often put in place to help public authorities in planning for the provision of public services such as sewerage, roads, public schools, health services, and public transportation. In addition, zoning regulations often serve to minimize negative externalities by segregating offending land uses. But regulations can affect the supply of affordable housing through various channels. Consumption standards such as minimum lot sizes can exclude low-income groups if they are set too high. Regulations such as density limits or height restrictions reduce affordability by reducing the amount of housing the land can accommodate, thus cutting supply. Urban growth boundaries (UGBs; sometimes called greenbelts) can raise housing prices by restricting the amount of developable land in a city. Alterman (2014) provides a survey of planning laws and urban development controls along with their implications for developing countries.

With some land-management regulations potentially having negative effects, it is natural to wonder why they would be imposed in the first place. As one explanation, Bertaud (2014) points to the predilection of urban planners for regulating the urban landscape rather than relying on market forces to guide land-use outcomes. Under this view, planners are well meaning but, in their zeal to shape cities, sometimes choose policies with unintended and unfavorable consequences for housing affordability and the welfare of urban residents. Regulations such as UGBs may also reflect the exercise of monopoly power by existing landowners, who can enrich themselves via policies that restrict housing supply and thus raise prices (see Brueckner, 1999 for a survey of the literature on urban “growth controls”). Finally, in developing countries, welfare-reducing urban regulations could also arise from a desire to limit rural–urban migration, mirroring the motivations explored by Feler and Henderson (2011).

21.5.2.1 Land-consumption standards

In the cities of developing countries, minimum land-consumption standards in the formal sector are often incompatible with the level of economic development and the incomes

of the poor. While general-purpose land-use regulations can improve the functioning of the formal land and housing markets, minimum-lot-size standards raise the effective house price to income ratio, making formal housing unaffordable for the poor. For example, the minimum plot size in Dar es Salaam is 500 m², and the authorities are discussing whether to raise it to 700 m² (Collier and Venables, 2013). In Nairobi, the minimum legal plot size is 1/16 of an acre, which is unaffordable for ordinary households. Such standards are often an overhang of colonial planning, notably the 1947 British Town and Country Planning Act, which is dominant in East Africa and Southern Africa. Regardless of the fact that African per capita incomes are less than 5% of British per capita incomes, many African regulatory authorities consider modernization as synonymous with the raising of standards from time to time.

With minimum-lot-size standards making formal housing unaffordable, poor households choose to live in informal settlements, either by purchasing informally subdivided land or houses built on such developments. However, the standards may also serve as newcomer taxes for potential migrants, and they may in fact reduce migration into the city, mirroring the effect of the denial of water connections analyzed by Feler and Henderson (2011). Lall et al. (2006b) empirically explore this question by using data from four rounds of the Brazilian census. They examine the implications of state and municipal decisions to waive federal subdivision regulations and reduce the minimum lot sizes on land set aside for developing low-income housing.

In 1979, the federal government in Brazil established national legislation (Federal Law 6766) for developing, approving, and registering urban land subdivisions (World Bank, 2006). Among the parameters are a minimum lot size of 125 m², with a minimum frontage of 5 m, and a compulsory donation of 35% of the development area for public uses and open space. Going against the national mandate, cities such as Recife, Belo Horizonte, Porto Alegre, and Belém have classified parts of their jurisdictions as Special Zones of Social Interest in order to regularize informal settlements and produce affordable housing for the poor. The Special Zones of Social Interest have flexible zoning regulations such as reduced minimum lot sizes (90 m² in Belém, 50 m² in Fortaleza, and 40 m² in Belo Horizonte) and variable frontage (World Bank, 2006). Lall et al. (2006b) find that relaxing minimum-lot-size regulations increases housing supply but also prompts higher population growth. This growth is in fact faster than the formal housing-supply response, leading to an increase in slum formation. Such an outcome suggests that policies that aim to reduce affordability barriers need to be accompanied by other steps to raise housing supply. In the absence of these measures, pro-poor land regulations may in fact exacerbate slum formation.

21.5.2.2 Density regulations

Density regulations, which are one of the tools used most frequently by urban planners, cap the quantity of housing that can be developed on a plot of land. Such a cap is usually

imposed by restricting the floor-space ratio (FAR) on the plot (sometimes called the floor-space index, or FSI). For example, if the FAR limit is 1, developers can only build a structure whose total floor space is less than or equal to the plot area. While in some cases it may be possible to build a one-story building on a plot that entirely covers it (therefore achieving an FAR of 1), developers will typically construct a building with a footprint smaller than the plot. For example, a developer could cover 25% of a plot and build a four-story building while still meeting the FAR limit of 1 ([World Bank, 2013](#)).

From the point of view of urban planners (see [Bertaud, 2004, 2014](#)), the “right” FAR limit in an area will depend on the existing spatial structure of the city, the street patterns and widths, the level of infrastructure (is there enough capacity to accommodate high densities?), and cultural and social factors (are skyscrapers acceptable?). In fact, there is considerable variation in FAR limits across the world’s cities, with values ranging from 1 to 25, as seen in [Table 21.2](#). [Table 21.2](#) shows that, except for São Paulo, most of the lowest FAR values are in India ([World Bank, 2013](#)). Other cities have much higher FAR limits, ranging from 3 for Paris to 25 for Singapore. Most of the cities with high FAR values tend to have substantial levels of infrastructure per acre.

An emerging body of evidence (see [Annez and Linn, 2010](#)) suggests that if an area’s FAR limit is set far below the level that would be chosen in the absence of regulation, the resulting reduction in housing supply will raise prices throughout the city, reducing housing affordability. In addition, the urban footprint will grow, pushing residences farther from job centers. [Bertaud and Brueckner \(2005\)](#) demonstrate these effects in a

Table 21.2 FAR limits in central business districts

City	FAR limit
São Paulo, Brazil	1
Mumbai, India	1.33
Chennai, India	1.5
Delhi, India	1.2–3.5
Amsterdam, Netherlands	1.9
Venice, Italy	2.4
Paris, France	3
Shanghai, China	8
Vancouver, Canada	8
San Francisco, United States	9
Chicago, United States	12
Hong Kong Special Administrative Region, China	12
Los Angeles, United States	13
New York, United States	15
Denver, United States	17
Tokyo, Japan	20
Singapore, Singapore	12–25

theoretical model, and they show that the resulting welfare loss for individual households can be measured by the increase in the commuting cost for the household at the edge of the city. Applying this result to the Indian city of Bengaluru (Bangalore), where FAR limits range between 1.75 and 3.25, the study predicts that the city radius would shrink from 12 to 8 km if the limits were removed. This shrinkage would reduce the commuting costs of edge households by 4% of their income, a gain that captures the welfare loss from the FAR limit. Rather than using a simulation exercise to reach such a conclusion, [Brueckner and Sridhar \(2012\)](#) rely on empirical estimates. They use a cross-sectional regression that relates the size of the urban footprints of Indian cities to their central FAR limits and other variables such as income and population. The results, which show that a higher FAR limit indeed shrinks the urban footprint, allow calculation of the resulting reduction in commuting distance for the edge household and thus the saving in commuting cost. This calculation shows that a unitary increase in a city's FAR limit (from an average of near 3 to 4) yields a commuting-cost saving of 0.7% of income for edge households, indicating a substantial welfare gain from a marginal relaxation in this type of land-use regulation.

While FAR limits reduce development densities and thus housing supply and prices, subdivision regulations can also have a density-reducing effect while raising housing prices through other cost-related channels. This class of interventions is well illustrated in the analyses of [Malpezzi and Mayo \(1997\)](#) and [Bertaud and Malpezzi \(2001\)](#), who study the case of Malaysia. They focus on regulations that require excessive road widths in newly developed areas (including provision of back alleys), excessive street setbacks for structures, and excessive requirements for community facilities in new developments. These authors also identify a less tangible cost-increasing factor, which grows out of the process for securing government approval of new projects. Long regulatory delays in the approval process raise the cost of development, as does uncertainty about the outcome of the process. [Malpezzi and Mayo \(1997\)](#) are able to generate a precise numerical estimate of the impact of these factors on the cost per dwelling. They focus on a representative dwelling with a delivery cost of about 28,000 Malaysian dollars, and they attribute 4000 Malaysian dollars of this amount to the effects of government regulations, a share of about 14%. In addition, [Bertaud and Malpezzi \(2001\)](#) find that if restrictions on construction and road patterns were eased so as to increase salable land to 55% of the developable land, developers' profits would double, providing a greater incentive to shift their activities to the lower-income market.

21.5.2.3 Urban growth boundaries

It is well known that UGBs can raise housing prices by restricting the amount of land available for development (see [Brueckner, 2001](#); [Cheshire and Sheppard, 2002](#)). UGBs are present in the developed world, but the urbanization process in some developing countries, notably in South Korea in past years, has been affected by their presence.

Because cities constrained by UGBs differ from non-UGB cities in a host of other ways (including the presence of additional land-use restrictions), it is difficult to use cross-sectional regression analysis to isolate the impact of a UGB on variables such as housing prices. Despite this limitation, researchers studying South Korea present persuasive evidence that the country's UGBs (greenbelts) contributed in the past to its high housing prices. For example, data presented by [Hannah et al. \(1993\)](#) show that an index of Korean housing prices grew by a factor of 10 between 1974 and 1989, while real GDP rose by a factor of 3.4. Although the effect of rising incomes on demand no doubt played a role in this price escalation, the economy also faced severe land constraints. Urban residential land in South Korea grew by only 65% between 1973 and 1988 even though the urban population rose by more than 100% over that period. With the greenbelt policy slowing the growth in the supply of residential land, and with population and income both rising rapidly, the conditions were ripe for a rapid escalation in Korean housing prices and falling affordability.

21.5.3 Investment in complementary infrastructure and services

For affordable housing to increase household welfare, the structure needs to be supported by complementary physical infrastructure and social services such as roads and transport services, drainage, street lighting, electricity, water, and sewerage, together with policing, schools, waste disposal, and health care. The benefits of such services make urban living more “affordable” in a sense that goes beyond housing costs. While the capital costs of some of these services could be covered by private property developers, supply is ultimately best undertaken publicly given that the services tend to involve coordination problems. Some of the services are network-based services that cannot be provided to each household individually, and some are services that, when supplied to separate households, create substantial externalities, as in the case of sanitation ([Collier, 2013](#)). Since such services will be underprovided and underconsumed relative to the social optimum in the absence of coordination efforts, public provision is desirable.

Empirical work on valuation of services relies on contingent valuation ([Cameron, 1992](#); [Whittington, 2002](#)), conjoint and discrete-choice analysis ([Earnhart, 2002](#)), and estimation of hedonic models ([Chattopadhyay, 1999](#); [Malpezzi, 2003](#); [Palmquist, 2005](#)). Several recent studies employing data from georeferenced household surveys extend these methods for valuing public services. [Lall and Lundberg \(2008\)](#) use a non-parametric approach to improve on the traditional hedonic method. They address the concerns about functional-form bias in hedonic valuation methods raised by [Malpezzi \(2003\)](#), using a variant of a three-stage estimation procedure developed in [Bajari and Kahn \(2005\)](#). Lall and Lundberg estimate a nonparametric hedonic function and recover the structural demand parameters using first-order conditions, following which they estimate the determinants of these demand parameters using a vector of household

characteristics. Estimating the model using georeferenced household survey data for Pune, India, they find that public services are valued relatively more by the poor than by higher-income households. The services examined (water, transport, education, safety) are worth roughly 29% of the rental value of the dwelling for the wealthy, but are valued at 67% of rent by the poor. While the list of services examined is not comprehensive, the results suggest that even untargeted, across-the-board investment in public services can be progressive. With access to services, the poor migrant's limited spending on housing buys more than just shelter, effectively increasing the affordability of urban living.

By focusing only on the gains from direct consumption of services, such estimates may undervalue the affordability benefits of improved service delivery when spillovers are present. [Anselin et al. \(2010\)](#) address this issue by taking advantage of a unique georeferenced household survey from Bengaluru (Bangalore), India. They carry out a hedonic analysis of housing values that explicitly accounts for spatial spillovers in provision of a water service. Building on the standard hedonic literature, they use a spatially explicit framework that allows measurement of both direct effects and externality spillovers from upgrades in water systems made by neighbors. The results show that standard nonspatial hedonic benefit estimates are around 30% lower than estimates that consider spatial spillovers. By raising benefits, such spillovers enlarge the housing/public-service bundle associated with a given rental payment, thus improving affordability.

21.6. CONCLUSION

This chapter has reviewed major strands of theoretical and empirical work on urbanization and housing provision in developing countries, focusing on the three elements highlighted in the chapter's title. The initial focus is on rural-rural migration, a main driver of city growth in the developing world. The discussion summarizes the state of empirical research on such migration and sketches theoretical work in the Harris-Todaro tradition, which depicts the urbanization equilibrium generated by the migration process. The chapter then turns to the issue of tenure insecurity in developing countries, focusing mainly on the economics of squatting from both theoretical and empirical perspectives. Squatting is partly a response to limited housing affordability in developing countries, which was the third topic considered in the chapter. Following a macro-oriented comparison of housing investment in low- and high-income countries, the discussion argues that misguided land-management policies in the developing world often limit housing affordability, while also pointing out that urban public services must be part of an affordable housing package.

More generally, drawing a distinction between urbanization, on the one hand, and city development, on the other, can be beneficial in understanding the experience of developing countries. These processes are distinct since much of urbanization takes place early in development, before countries have reached middle incomes. In contrast, city

development in the form of investments in durable structures (notably housing) and infrastructure accelerates later as countries approach middle incomes. This view implies that urbanization is a messy process, where surges in population density precede the fixed investments needed to properly house, transport, and service the urban population. To manage this messy process, there is a need to identify beneficial policies or policy changes that can be implemented by countries with limited administrative and financial capabilities. Three instances are particularly relevant.

First, land management must be improved by eliminating those policies that inadvertently raise the cost of housing, as discussed in [Section 21.5](#). By making shelter more affordable, such policy changes will allow cities in developing countries to do a better job of housing their burgeoning populations. Second, the provision of urban infrastructure must be managed more effectively so as to improve the delivery of public services. Provision of some infrastructure (streets, basic drainage) can be left to private developers, while infrastructure with a broader spatial scope (local road networks, schools) should be the responsibility of city wards, with citywide investments (transport, water, sewerage) or regional infrastructure (electricity) handled at even higher levels. Devolution of responsibilities to the lowest possible level may be beneficial when central governments are weak. In making transportation investments, governments should recognize that, by easing access to jobs, such investment can improve the functioning of urban labor markets. Moreover, in making all types of infrastructure investments, governments should take into account the attractive power of public services, which tend to spur rural–urban migration.

Third, housing investment can be spurred by capital-market improvements, which should also be a policy goal. As mentioned earlier, since tenure insecurity inhibits the development of mortgage markets, steps to accelerate the process of conveying titles to residents in formal housing could spur development of mortgage markets, unleashing a mortgage-financed surge in housing investment. Granting titles to squatters and other informal residents would have a parallel effect. In addition, following the lead of the United States and other developed countries, the developed world could marshal funds for housing investment by fostering the securitization of mortgage debt once local mortgage markets gain sufficient breadth.

The chapter points to a number of possible avenues for future research. On the empirical aspects of rural–urban migration, the recent work of [Giulietti et al. \(2014\)](#), which applies social-network theory to understand better the interpersonal forces that spur migration, seems to be a fruitful area for further research. This social-network approach could be applied to individual-level data from countries other than China, on which these authors focused. The empirical underpinnings of the migration models considered in [Section 21.3](#) could also be explored further. For example, the models predict equalization of rural income and the income net of the commuting cost of city workers living at the urban fringe, a relationship that could be explored empirically (in principle, at least) if cross-country data were available. Even by itself, the urban net income measure would be useful since it appropriately captures urban living standards in a developing country

without the need for data on housing prices. Whether the migration process produces cities that are close to optimal in size is another question that could be addressed. Researchers could follow the method in Au and Henderson's (2006) important China study, focusing on other countries.

A number of different avenues for work on tenure insecurity are also apparent. Research on squatting would greatly benefit from international data that track squatter populations as well as eviction events across developing countries. Population estimates could be based on measurement of the land areas of squatter settlements, a task that could make use of satellite data supplemented by local expertise regarding the boundaries of the settlements. Although collection of eviction data would require monitoring of press coverage in a large number of cities around the world, the task is increasingly feasible in today's information age. Another squatting-related empirical exercise would be a test of the notion that squatter settlements squeeze the formal housing market. Empirically, the goal would be to test the prediction that formal housing prices are higher in cities with a large share of squatter households (requiring better population figures along with price data). While less general in scope, country case studies such as those of Selod and Tobin (2013) and Kim (2004) provide highly useful insights into the extent of tenure insecurity, making further studies worthwhile.

Research on housing affordability could be directed toward generating quantitative measures of the affordability gains from relaxing counterproductive land-use policies. Malpezzi and Mayo (1997) and Bertaud and Malpezzi (2001) show how to produce such measures from detailed information on building regulations, while Brueckner and Sridhar (2012) measure the gain from looser building-height restrictions using a more aggregative approach. Such quantitative information provides crucial leverage for the World Bank and other international agencies in persuading governments to reform regulation of land use, and more studies of this type will assist them in this task.

APPENDIX

With variable q , the condition $N = \pi \bar{x}^2$ is replaced by $\int_0^{\bar{x}} [2\pi x/q(x, N)] dx - N \equiv \Phi - N = 0$. Total differentiation of this condition yields $\partial \bar{x} / \partial N = (1 - \Phi)(\bar{q}/2\pi \bar{x})$, where $\bar{q} \equiv q(\bar{x}, N)$. Differentiating $\text{ACC} \equiv \int_0^{\bar{x}} [2\pi x/q(x, N)] t x dx$ with respect to N then yields

$$\begin{aligned} \frac{\partial \text{ACC}}{\partial N} &= \frac{\partial \text{ACC}}{\partial \bar{x}} \frac{\partial \bar{x}}{\partial N} - \int_0^{\bar{x}} \frac{2\pi x}{q^2} \frac{\partial q}{\partial N} t x dx \\ &= \frac{2\pi \bar{x}^2 t}{\bar{q}} (1 - \Phi) \frac{\bar{q}}{2\pi \bar{x}} - \int_0^{\bar{x}} \frac{2\pi x}{q^2} \frac{\partial q}{\partial N} t x dx = t \bar{x} + \int_0^{\bar{x}} \frac{2\pi x}{q^2} \frac{\partial q}{\partial N} t (\bar{x} - x) dx \\ &\equiv t \bar{x} - \Omega(N), \end{aligned} \tag{A.1}$$

where $\Omega(N) > 0$ equals minus the negative integral in the penultimate line, which represents the reduction in ACC due to higher densities (lower values of q). Subtracting $\partial \text{ACC} / \partial N$ from the increase in urban income then yields (21.13).

ACKNOWLEDGMENTS

We thank the editors, Susan Wachter, and Maisy Wong for helpful comments. Any shortcomings in the chapter, however, are our responsibility.

REFERENCES

- Ades, A.F., Glaeser, E.L., 1995. Trade and circuses: explaining urban giants. *Q. J. Econ.* 110, 195–227.
- Alonso, W., 1964. *Location and Land Use*. Harvard University Press, Cambridge.
- Alterman, R., 2014. Planning laws, development controls, and social equity: lessons for developing countries. *World Bank Law Rev.* 5, 329–350.
- Anderson, K., Pomfret, R., 2005. Spatial inequality and development in central Asia. Unpublished paper—World Institute for Development Economic Research (UNU-WIDER).
- Annez, P.C., Linn, J.F., 2010. An agenda for research on urbanization in developing countries: a summary of findings from a scoping exercise. World Bank Policy Research Working paper No. 5476.
- Anselin, L., Lozano-Gracia, N., Deichmann, U., Lall, S., 2010. Valuing access to water: a spatial hedonic approach, with an application to Bangalore. *Indian Spat. Econ. Anal.* 5, 161–179.
- Arnott, R., 1979. Optimal city size in a spatial economy. *J. Urban Econ.* 6, 65–89.
- Au, C.-C., Henderson, J.V., 2006. Are Chinese cities too small? *Rev. Econ. Stud.* 73, 549–576.
- Bajari, P., Kahn, M.E., 2005. Estimating housing demand with an application to explaining racial segregation in cities. *J. Busin. Econ. Stat.* 23, 20–33.
- Barrios, S., Bertinelli, L., Strobl, E., 2006. Climatic change and rural-urban migration: the case of sub-Saharan Africa. *J. Urban Econ.* 60, 357–371.
- Bertaud, A., 2004. Mumbai FSI/FAR conundrum: the perfect storm: the four factors restricting the construction of new floor space in Mumbai. http://alain-bertaud.com/AB_Files/AB_Mumbai_FSI_conundrum.pdf.
- Bertaud, A., 2014. The formation of urban spatial structures: markets vs. design. Working paper No. 7, Marron Institute, New York University.
- Bertaud, A., Brueckner, J.K., 2005. Analyzing building-height restrictions: predicted impacts and welfare costs. *Reg. Sci. Urban Econ.* 35, 109–125.
- Bertaud, A., Malpezzi, S., 2001. Measuring the costs and benefits of urban land use regulation: a simple model with an application to Malaysia. *J. Hous. Econ.* 10, 393–418.
- Brueckner, J.K., 1990. Analyzing third world urbanization: a model with empirical evidence. *Econ. Dev. Cult. Chang.* 38, 587–610.
- Brueckner, J.K., 1999. Modeling urban growth controls. In: Panagariya, A., Portney, P., Schwab, R.M. (Eds.), *Environmental and Public Economics: Essays in Honor of Wallace E. Oates*. Edward Elgar Publishers, London, pp. 151–168.
- Brueckner, J.K., 2001. Urban sprawl: lessons from urban economics. In: Gale, W.G., Pack, J.R. (Eds.), *Brookings-Wharton Papers on Urban Affairs*. Brookings Institution, Washington, DC, pp. 65–89.
- Brueckner, J.K., 2013a. Slums in developing countries: new evidence for Indonesia. *J. Hous. Econ.* 22, 278–290.
- Brueckner, J.K., 2013b. Urban squatting with rent-seeking organizers. *Reg. Sci. Urban Econ.* 43, 561–569.
- Brueckner, J.K., Kim, H.-A., 2001. Land markets in the Harris-Todaro model: a new factor equilibrating rural-urban migration. *J. Reg. Sci.* 41, 507–520.
- Brueckner, J.K., Selod, H., 2009. A theory of urban squatting and land-tenure formalization in developing countries. *Am. Econ. J. Econ. Pol.* 1, 28–51.

- Brueckner, J.K., Sridhar, K.S., 2012. Measuring welfare gains from relaxation of land-use restrictions: the case of india's building-height limits. *Reg. Sci. Urban Econ.* 42, 1061–1067.
- Brueckner, J.K., Zenou, Y., 1999. Harris-Todaro models with a land market. *Reg. Sci. Urban Econ.* 29, 317–339.
- Bryan, G., Chowdhury, S., Mobarak, A.M., 2014. Under-investment in a profitable technology: the case of seasonal migration in Bangladesh. *Econometrica* 82, 1671–1748.
- Bryceson, D.F., 1999. Sub-Saharan Africa betwixt and between: rural livelihood practices and policies. Working paper 43/1999, Leiden, University of.
- Burns, L.S., Grebler, L., 1976. Resource allocation to housing investment: a comparative international study. *Econ. Dev. Cult. Chang.* 25, 95–121.
- Cameron, T., 1992. Combining contingent valuation and travel cost data for the valuation of non-market goods. *Land Econ.* 68, 302–317.
- Chattopadhyay, S., 1999. Estimating the demand for air quality: new evidence based on the Chicago housing market. *Land Econ.* 75, 22–38.
- Cheshire, P., Sheppard, S., 2002. The welfare economics of land use planning. *J. Urban Econ.* 52, 242–269.
- Collier, P., 2013. Building African cities that work. Working paper, Oxford University.
- Collier, P., Venables, A.J., 2013. Housing and urbanization in Africa: unleashing a formal market process. In: Glaeser, E., Joshi-Ghani, A. (Eds.), *Rethinking Cities*. World Bank, Washington, DC (forthcoming).
- Daniere, A.G., 1994. Estimating willingness-to-pay for housing attributes an application to Cairo and Manila. *Reg. Sci. Urban Econ.* 24, 577–599.
- Dasgupta, B., Lall, S.V., Lozano, N., 2014. Urbanization and housing investment. Unpublished paper. World Bank.
- Davis, J.C., Henderson, J.V., 2003. Evidence on the political economy of the urbanization process. *J. Urban Econ.* 53, 98–125.
- Deshingkar, P., Grimm, S., 2004. Internal migration and development: a global perspective. IOM Migr. Res. Ser. 19, Geneva.
- De Soto, H., 2000. *The Mystery of Capital: Why Capitalism Triumphs in the West and Fails Everywhere Else*. Basic Books, New York, NY.
- Di Tella, R., Galiani, S., Schargodsky, E., 2007. The formation of beliefs: evidence from the allocation of land titles to squatters. *Q. J. Econ.* 122, 209–241.
- Dudwick, N., Hull, K., Katayama, R., Shilpi, F., Simler, K., 2011. *From Farm to Firm: Rural-Urban Transition in Developing Countries*. World Bank, Washington, DC.
- Earnhart, D., 2002. Combining revealed and stated data to examine housing decisions using discrete choice analysis. *J. Urban Econ.* 51, 143–169.
- Feler, L., Henderson, J.V., 2011. Exclusionary policies in urban development: under-servicing migrant households in Brazilian cities. *J. Urban Econ.* 69, 253–272.
- Ferre, C., 2009. Is internal migration bad for receiving urban centers? Evidence from Brazil (1995–2000). Unpublished paper. World Bank.
- Field, E., 2005. Property rights and investment in urban slums. *J. Eur. Econ. Assoc. Pap. Proc.* 3, 279–290.
- Field, E., 2007. Entitled to work: urban property rights and labor supply in Peru. *Q. J. Econ.* 122, 1561–1602.
- Fields, G., 1982. Place-to-place migration in Colombia. *Econ. Dev. Cult. Chang.* 30, 539–558.
- Flatters, F., Henderson, V., Mieszkowski, P., 1974. Public goods, efficiency, and regional fiscal equalization. *J. Publ. Econ.* 3, 99–112.
- Follain, J., Lim, G.-C., Renaud, B., 1982. Housing crowding in developing countries and willingness to pay for additional space: the case of Korea. *J. Dev. Econ.* 11, 249–272.
- Follain, J.R., Jimenez, E., 1985. The demand for housing characteristics in developing countries. *Urban Stud.* 22, 421–432.
- Friedman, J., Jimenez, E., Mayo, S., 1988. The demand for tenure security in developing countries. *J. Dev. Econ.* 29, 185–198.
- Galiani, S., Schargodsky, E., 2010. Property rights for the poor: effects of land titling. *J. Publ. Econ.* 94, 700–729.

- Giulietti, C., Wahba, J., Zenou, Y., 2014. Strong versus weak ties in migration. Unpublished paper, Stockholm University.
- Granovetter, M.S., 1973. The strength of weak ties. *Am. J. Soc.* 78, 1360–1380.
- Gross, D.J., 1988. Estimating willingness to pay for housing characteristics: an application of the Ellickson bid rent model. *J. Urban Econ.* 24, 95–112.
- Hammam, S., 2013. Housing matters. In: Glaeser, E., Joshi-Ghani, A. (Eds.), *Rethinking Cities*. World Bank, Washington, DC (forthcoming).
- Hannah, L., Kim, K.-H., Mills, E.S., 1993. Land use controls and housing prices in Korea. *Urban Stud.* 30, 147–156.
- Hardoy, J., Satterthwaite, D., 1989. *Squatter Citizen: Life in the Urban Third World*. Earthscan, London.
- Harris, J.R., Todaro, M.P., 1970. Migration, unemployment and development: a two-sector analysis. *Am. Econ. Rev.* 60, 126–142.
- Henderson, J.V., Wang, H.G., 2005. Aspects of the rural-urban transition of countries. *J. Econ. Geogr.* 5, 23–42.
- Henderson, J.V., Wang, H.G., 2007. Urbanization and city growth: the role of institutions. *Reg. Sci. Urban Econ.* 37, 283–313.
- Henry, S., Boyle, P., Lambin, E.F., 2003. Modelling inter-provincial migration in Burkina Faso, West Africa: the role of socio-demographic and environmental factors. *Appl. Geogr.* 23, 115–136.
- Hidalgo, F.D., Naidu, S., Nichter, S., Richardson, N., 2010. Economic determinants of land invasions. *Rev. Econ. Stat.* 92, 505–523.
- Hoy, M., Jimenez, E., 1991. Squatters' rights and urban development: an economic perspective. *Economica* 58, 79–92.
- Hunter, L.M., Murray, S., Riosmena, F., 2011. Climatic variability and U.S. migration from rural Mexico. IBS Working paper.
- Iliffe, J., 1995. *Africans: The History of a Continent*. Cambridge University Press, Cambridge.
- Jimenez, E., 1984. Tenure security and urban squatting. *Rev. Econ. Stat.* 66, 556–567.
- Jimenez, E., 1985. Urban squatting and community organization in developing countries. *J. Publ. Econ.* 27, 69–92.
- Kapoor, M., le Blanc, D., 2008. Measuring risk on investment in informal (illegal) housing: theory and evidence from Pune. India. *Reg. Sci. Urban Econ.* 38, 311–329.
- Kiernan, P., 2014. Rio evicts squatters, sparking riot. *Wall Street J* A9, April 12.
- Kim, A., 2004. A market without the 'right' property rights. *Econ. Trans.* 12, 275–305.
- Lall, S.V., Lundberg, M., 2008. What are public services worth, and to whom? Non-parametric estimation of capitalization in Pune. *J. Hous. Econ.* 17, 34–64.
- Lall, S.V., Selod, H., Shalizi, Z., 2006a. Rural-urban migration in developing countries: a survey of theoretical predictions and empirical findings. World Bank Policy Research Working paper No. 3915.
- Lall, S.V., Wang, H.G., da Mata, D., 2006b. Do urban land regulations influence slum formation? Evidence from Brazilian Cities. Unpublished paper, World Bank.
- Lall, S., Lundberg, M., Shalizi, Z., 2008. Implications of alternate policies on welfare of slum dwellers: evidence from Pune. India. *J. Urban Econ.* 63, 56–73.
- Lall, S.V., Timmins, C., Yu, S., 2009. Connecting lagging and leading regions: the role of labor mobility. World Bank Policy Research, Working paper No. 4843.
- Lanjouw, J., Levy, P., 2002. Untitled: a study of formal and informal property rights in urban Ecuador. *Econ. J.* 112, 986–1019.
- Lim, G.-C., Follain, J., Renaud, B., 1984. Economics of residential crowding in developing countries. *J. Urban Econ.* 16, 173–186.
- Lozano, N., Young, C., 2014. Housing demand in urban africa. Unpublished paper. World Bank.
- Lozano-Gracia, N., Young, C., Lall, S.V., Vishwanath, T., 2013. Leveraging land to enable urban transformation: lessons from global experience. World Bank Policy Research Working paper 6312.
- Lucas Jr., R.E., 2004. Life earnings and rural-urban migration. *J. Polit. Econ.* 112, S29–S59.
- Malpezzi, S., 2003. Hedonic pricing models: a selective and applied review. In: O'Sullivan, A., Gibb, K. (Eds.), *Housing Economics and Public Policy*. Blackwell Science, Oxford, pp. 67–85.
- Malpezzi, S., 2012. Global perspectives on housing markets. Unpublished paper, University of Wisconsin.

- Malpezzi, S., Mayo, S.K., 1997. Getting housing incentives right: a case study of the effects of regulation, taxes, and subsidies on housing supply in Malaysia. *Land Econ.* 7, 372–391.
- Marre, A.W., 2009. Rural out-migration, income, and poverty: are those who move truly better off? USDA-Economic Research Center Working paper.
- Marx, B., Stoker, T., Suri, T., 2013. The economics of slums in the developing world. *J. Econ. Perspect.* 27, 187–210.
- Mills, E.S., 1967. An aggregative model of resource allocation in a metropolitan area. *Am. Econ. Rev.* 57, 197–210.
- Moomaw, R.L., Shatter, A.M., 1996. Urbanization and economic development: a bias toward large cities? *J. Urban Econ.* 40, 13–37.
- Mora, J., Taylor, E., 2007. Determinants of migration, destination, and sector choice: disentangling individual, household, and community effects. In: Ozden, C., Schiff, M. (Eds.), *International Migration, Remittances, and Brain Drain*. Palgrave Macmillan, New York, NY, pp. 21–52.
- Mourmouras, A., Rangazas, P., 2013. Efficient urban bias. *J. Econ. Geogr.* 13, 451–471.
- Munshi, K., Rosenzweig, M., 2013. Networks and misallocation: insurance, migration, and the rural-urban wage gap. Unpublished paper, Yale University.
- Muth, R.F., 1969. *Cities and Housing*. University of Chicago Press, Chicago, IL.
- Mutlu, S., 1989. Urban concentration and primacy revisited: an analysis and some policy conclusions. *Econ. Dev. Cult. Chang.* 37, 611–639.
- Palmquist, R.B., 2005. Property value models. In: Máler, K., Vincent, J. (Eds.), *Handbook of Environmental Economics*, vol. 2. North-Holland, Amsterdam, pp. 763–813.
- Poelhekke, S., 2011. Urban growth and uninsured rural risk: booming towns in bust times. *J. Dev. Econ.* 96, 461–475.
- Quigley, J.M., 1982. Nonlinear budget constraints and consumer demand: an application to public programs for residential housing. *J. Urban Econ.* 12, 177–201.
- Regmi, A., Deepak, M.S., Seale Jr., J.L., Bernstein, J., 2001. Cross-country analysis of food consumption patterns. In: USDA, E.R.S. (Ed.), *Changing Structure of Global Food Consumption and Trade*. United States Department of Agriculture, Washington, DC, pp. 14–22.
- Reuveny, R., 2007. Climate change-induced migration and violent conflict. *Pol. Geogr.* 26, 656–673.
- Rosen, K.T., Resnick, M., 1978. The size distribution of cities: an examination of the Pareto law and primacy. *J. Urban Econ.* 8, 165–186.
- Sahn, D., Stifel, D., 2003. Urban-rural inequality in living standards in Africa. *J. Afr. Econ.* 12, 564–597.
- Schultz, T.P., 1982. Lifetime migration within educational strata in Venezuela: estimates of a logistic model. *Econ. Dev. Cult. Chang.* 30, 559–593.
- Schwartz, A., 1973. Interpreting the effect of distance on migration. *J. Pol. Econ.* 81, 1153–1169.
- Selod, H., Tobin, L., 2013. City structure and informal property rights in West Africa: theory and evidence. Unpublished paper. World Bank.
- Shah, N., 2014. Squatting on government land. *J. Reg. Sci.* 54, 114–136.
- Takeuchi, A., Cropper, M., Bento, A., 2008. Measuring the welfare effects of slum improvement programs: the case of Mumbai. *J. Urban Econ.* 64, 65–84.
- Todaro, M.P., 1976. *Internal Migration in Developing Countries: A Review of Theory, Evidence, Methodology and Research Priorities*. International Labour Office, Geneva.
- Turnbull, G., 2008. Squatting, eviction, and development. *Reg. Sci. Urban Econ.* 38, 1–15.
- United Nations Habitat, 2012. *State of the World's Cities*. United Nations Human Settlement Programme, London.
- Wandschneider, T., Mishra, P., 2003. The role of small rural towns in Bolangir District, India: a village-level perspective. NRI Report No. 2750, DFID-World Bank Collaborative Research Project on the Rural Non-Farm Economy and Livelihood Enhancement.
- Whalley, J., Zhang, S., 2004. Inequality change in China and (hukou) labour mobility restrictions. National Bureau of Economic Research Working paper No. 10683.
- Wheaton, W., Shishido, H., 1981. Urban concentration, agglomeration economies and the level of economic development. *Econ. Dev. Cult. Chang.* 100, 17–30.
- Wheaton, W.C., 1974. A comparative static analysis of urban spatial structure. *J. Econ. Theory* 9, 223–237.

- Whittington, D., 2002. Improving the performance of contingent valuation studies in developing countries. *Environ. Resour. Econ.* 22, 23–367.
- World Bank, 2006. Brazil—inputs for a strategy for cities: a contribution with a focus on cities and municipalities. **Report 35749-BR**. World Bank, Washington, DC.
- World Bank, 2007. Dhaka: improving living conditions for the urban poor. World Bank, Bangladesh Development Series Paper No. 17.
- World Bank, 2008. World Development Report 2009: Reshaping Economic Geography. World Bank, Washington, DC.
- World Bank, 2009. World Development Report. World Bank, Washington, DC.
- World Bank, 2010. Connecting People to Prosperity: Reshaping Economic Geography in Sri Lanka. World Bank, Colombo.
- World Bank, 2013. Planning, Connecting, and Financing Cities Now: Priorities for City Leaders. World Bank, Washington, DC.
- Young, A., 2013. Inequality, the urban-rural gap, and migration. *Q. J. Econ.* 128, 1727–1785.

This page intentionally left blank

CHAPTER 22

The Geography of Development Within Countries

Klaus Desmet^{*}, J. Vernon Henderson[†]

^{*}Department of Economics, Southern Methodist University, Dallas, TX, USA

[†]Department of Geography, London School of Economics, London, UK

Contents

22.1. Introduction	1458
22.2. Development and the Aggregate Spatial Distribution	1459
22.2.1 Development: Urban versus rural	1459
22.2.1.1 <i>Industrialization and urbanization</i>	1460
22.2.1.2 <i>Rural–urban migration and the transition to modern growth</i>	1461
22.2.1.3 <i>Rural–urban migration without industrialization</i>	1462
22.2.2 Development: Continuum of locations	1463
22.2.2.1 <i>Facts</i>	1464
22.2.2.2 <i>Theory</i>	1468
22.3. Development, Space, and Industries	1475
22.3.1 Manufacturing versus services	1476
22.3.2 Life cycle of industries and spatial distribution	1477
22.3.3 Ruralization versus suburbanization	1481
22.3.4 The cost of remoteness	1481
22.4. The Urban Sector	1482
22.4.1 Production patterns in the urban hierarchy	1483
22.4.1.1 <i>Facts</i>	1483
22.4.1.2 <i>Modeling the urban hierarchy</i>	1487
22.4.2 Dynamics in the urban hierarchy	1499
22.4.2.1 <i>Facts and concepts concerning the size distribution of cities</i>	1499
22.4.2.2 <i>Churning and movement of industries across the urban hierarchy</i>	1501
22.4.3 Policies affecting the spatial allocation of resources	1506
22.4.3.1 <i>Transport investments and technological change</i>	1508
22.4.3.2 <i>Urban and political city bias</i>	1508
22.5. Concluding Remarks	1512
References	1513

Abstract

This chapter describes how the spatial distribution of economic activity changes as economies develop and grow. We start with the relation between development and rural–urban migration. Moving beyond the coarse rural–urban distinction, we then focus on the continuum of locations in an economy and describe how the patterns of convergence and divergence change with development. As we discuss,

these spatial dynamics often mask important differences across sectors. We then turn our attention to the right tail of the distribution, the urban sector. We analyze how the urban hierarchy has changed over time in developed countries and more recently in developing countries. The chapter reviews both the empirical evidence and the theoretical models that can account for what we observe in the data. When discussing the stylized facts on geography and development, we draw on empirical evidence from both the historical evolution of today's developed economies and comparisons between today's developed and developing economies.

Keywords

Geography, Development, Space, Growth, City-size distribution, Spatial distribution of economic activity, Developed countries, Developing countries, Urban hierarchy, Industrialization and urbanization

JEL Classification Codes

R1, R11, R12, O18

22.1. INTRODUCTION

As economies grow and develop, the spatial distribution of the population, employment, and production changes. Probably the most prominent feature of this spatial transformation is increased urbanization. Between 1950 and 2009, the world's urban population more than quadrupled from 732 million to 3.4 billion, as the world moved from being under 30% urbanized to over 50% urbanized. Understanding the patterns of this rapid transformation is of paramount importance to policy makers. More than 80% of governments are concerned about the geographic distribution of people, and nearly 70% of them have implemented policies to reduce internal migration ([United Nations, 2010](#)). The goal of this chapter is to review what we know about the spatial distribution of economic activity and development. An important point we will make is that this spatial transformation can be viewed at different spatial scales and through different lenses. Which one is more useful will largely depend on the issue of interest.

One traditional divide is to contrast rural and urban areas, but that fails to capture the full richness of a country's spatial transformation. Rather than splitting up locations into two types (urban or rural), one often finds it useful to think of locations as a continuum, going from more rural (smaller and/or less dense) to more urban (larger and/or denser). The distribution of the population and economic activity along that continuum changes radically with development, and these changes mark how we view the overall geography of a country. What happens with aggregate employment and production often masks interesting differences across sectors. Manufacturing and services have exhibited very different spatial growth patterns over time.

Once a country becomes more urbanized, these changes and the spatial distribution are often viewed through a narrower lens that focuses on the urban sector. Within the urban sector there is enormous heterogeneity across the hierarchy of cities, and the

transformation of activities differs across that hierarchy. Finally, we note that while much of what we see is driven by market forces, the role of government in economies has grown. As a result, in today's developing countries, economic policies can have a strong effect on both the location and the concentration of economic activities.

This chapter reviews the models and evidence that characterize these processes. [Section 22.2](#) starts by looking at the urban–rural divide and then focuses on the continuum. It analyzes population and income convergence versus divergence and the reshaping of the location patterns of people and economic activity, especially in today's richer countries as they developed through the nineteenth century into the twentieth century. Another issue of interest that we discuss is the link between an economy's overall spatial structure and its aggregate growth. [Section 22.3](#) also focuses on the continuum, but takes a sectoral approach by looking at the structural transformation of economic activities as a country develops and matures. The distribution of economic activity differs across sectors, and these differences change over time as countries develop. [Section 22.4](#) looks at the urban sector, with particular attention on the urban hierarchy. It explores aspects of the transformation of the urban sector over the last 100 years in more developed countries and the more recent, rapid changes in developing countries. [Section 22.4](#) also discusses the key issue of how government policies in today's developing countries affect the transformation and the concentration of economic activities.

When discussing how the spatial distribution of economic activity changes with development, we draw on evidence both from comparing today's developed and developing economies and from analyzing the long-run evolution of today's developed countries. Although using historical evidence from today's developed countries to explain the spatial patterns of present-day developing countries is useful, this should be done with care. For example, because of trade and comparative advantage, the role of the structural transformation from agriculture to manufacturing in explaining urbanization in today's developing countries may be different from its role in nineteenth century Europe.

22.2. DEVELOPMENT AND THE AGGREGATE SPATIAL DISTRIBUTION

We start by discussing models of rural–urban migration. This coarse-grained look at the shift from the rural to the urban sector that occurs with development is the typical approach used by development economists. We cover recent developments to this paradigm that originally dates back to [Lewis \(1954\)](#). Then we turn to the perspective of a continuum which covers the national geography at a finer spatial scale.

22.2.1 Development: Urban versus rural

The link between urbanization and development has been emphasized both in the context of the transition from Malthusian to modern growth and in the work on rural–urban migration in developing countries. Much of the literature has emphasized the link between

development, industrialization, and urbanization. However, in light of the recent experience of Africa and the Middle East, urbanization and industrialization may not always go hand in hand, especially for countries whose incomes are heavily resource dependent.

22.2.1.1 Industrialization and urbanization

While the literature on the transition to modern growth is extensive, most of the competing models aim to capture the gradual transition from an agricultural-based rural economy to an industrial-based urban economy. In a context where incomes are growing, most articles generate this result by assuming an income elasticity of less than 1 for food items, leading to an increasing share of expenditure on urban goods. A simple way of modeling this is to introduce a subsistence constraint into standard Cobb–Douglas preferences, which yields a Stone–Geary utility function:

$$U(c_a, c_m) = (c_a - \bar{c}_a)^{1-\alpha} c_m^\alpha, \quad (22.1)$$

where c_a is agricultural consumption, c_m is manufacturing consumption, and $\bar{c}_a$ is the agricultural subsistence constraint. These preferences have been used in many models of industrialization (see, e.g., [Caselli and Coleman, 2001](#); [Desmet and Parente, 2012](#)). Such a setup creates a direct link between income *per capita*, industrialization, and urbanization, in as far as the industrial sector is less land intensive and more urbanized than the agricultural sector.

Another way of generating industrialization is by having an elasticity of substitution between agriculture and industry of less than 1:

$$U(c_a, c_m) = \left(\alpha_a c_a^{\frac{\sigma-1}{\sigma}} + \alpha_m c_m^{\frac{\sigma-1}{\sigma}} \right)^{\frac{\sigma}{\sigma-1}}, \quad (22.2)$$

where $\sigma < 1$. This approach to the structural transformation, taken by [Ngai and Pissarides \(2007\)](#) and [Desmet and Rossi-Hansberg \(2014a\)](#), implies that employment will shift out of agriculture into industry if agricultural productivity growth is higher.

Independently of whether we assume (22.1) or (22.2), an “agricultural revolution” must have preceded the industrial revolution. This idea is emphasized in the work by [Nurkse \(1953\)](#), [Rostow \(1960\)](#), [Schultz \(1968\)](#), and [Diamond \(1997\)](#) who argue that high agricultural productivity was a precondition for industrial takeoff. Consistent with this, [Allen \(2004\)](#) finds that output per worker in English agriculture doubled between 1600 and 1750, ahead of the industrial revolution. Greater agricultural efficiency allowed the economy to overcome the “food problem” and created a surplus of workers who could then engage in other activities, such as manufacturing. In modern developing countries, such as India, the Green Revolution has played a similar role. Work by [Gollin et al. \(2007\)](#) shows in a quantitative model that differences in agricultural total factor productivity (TFP) are key in explaining the differential timing of takeoff across countries. Note, however, that this positive link between agricultural productivity

and industrialization may be reversed when we allow for trade. As shown by Matsuyama (1992), in an open economy higher agricultural productivity may lock in a comparative advantage in that sector, thus delaying industrialization.

In most models of the industrial revolution and the transition to modern growth, the link to space and urban–rural migration is indirect. It is only in as far as we equate agriculture with rural and industry with urban that we get clear implications for the changing spatial distribution of economic activity. In some models, the transition from agriculture to manufacturing is implicit (Galor and Weil, 2000; Lucas, 2004), whereas in others it is explicit (Hansen and Prescott, 2002; Tamura, 2002; Doepke, 2004; Galor et al., 2009; Desmet and Parente, 2012). But in most of this literature, the focus is not on rural–urban migration *per se*. There are some exceptions though, such as Lucas (2004) and Henderson and Wang (2005), which we discuss in the next paragraphs.

22.2.1.2 Rural–urban migration and the transition to modern growth

Lucas (2004) proposes a model of infinitely lived dynasties to analyze the link between the structural transformation, urban–rural migration, and the shift from a traditional technology (with no growth) to a modern technology (with unbounded growth). In the rural sector, human capital is useless, whereas in the urban sector it increases productivity. Human capital accumulation depends on the time invested and on the human capital frontier.

The Lucas (2004) model captures some of the stylized facts of rural–urban migration. First, as the economy develops, people move gradually from the rural sector to the urban sector. Over time, as the human capital frontier moves out, it becomes less costly to accumulate human capital, making cities more attractive. The human capital externality—the fact that cities are good places to accumulate human capital—is key for this result. Second, recent arrivals do not work and instead spend their time improving their human capital. This is akin to the Harris and Todaro (1970) model where many of the recent arrivals are unemployed. The difference here is that unemployment is voluntary. Third, because the representative agent is a family, when migrants first arrive in the city, they are subsidized by the ones that stayed behind, and they later reimburse the rural part of the family through remittances.

In contrast to Lucas (2004), where there is only one consumption good, in Henderson and Wang (2005), the urban and rural sectors produce different goods. There is accumulation of human capital fueling growth in both sectors, although human capital externalities in the urban sector are posited to be greater. Demand for food is completely income inelastic. As human capital accumulates, and people become richer, there is a shift of the population and production out of the farm/rural sector into the urban/industrial sector and development. This leads to increased urbanization, as existing cities grow and new cities arise.

The fact that urbanization and industrialization often tend to go hand in hand (in many countries and models) raises the question of which one drives which. There are

good reasons to believe that causality runs both ways. In fact, the model of [Lucas \(2004\)](#) and many other models of the structural transformation feature such circular causality. Empirically, while urbanization certainly helps industrialization, it is not necessarily the case that industrialization started in the large urban centers. In the historical case of the US Northeast, for example, [Williamson \(1965\)](#) shows that the degree of urbanization exhibited convergence during the period of industrialization, suggesting that industrialization was particularly strong in the least urbanized places. The same was not true though for the United States as a whole, where there was divergence in urbanization during the nineteenth century. As in the US Northeast, also in England, many of the “hot spots” of the industrial revolution were initially relatively small towns that grew into large cities. Liverpool’s population, for example, multiplied by more than 60 times over the course of a century and a half, from 6000 in 1700 to 376,000 in 1850 ([Bairoch et al., 1988](#)). For developing countries, the issue is in part related to the adaptation of technologies: their agricultural and industrial revolutions often involve the simultaneous importing of world technologies in both sectors. In many developing countries, the rural sector has had fairly high levels of nonfarm activity and an important presence of traditional industries for decades.

22.2.1.3 Rural–urban migration without industrialization

Although urbanization and industrialization often go together, this is not always the case. [Gyourko et al. \(2013\)](#) document that urbanization in Africa (and the Middle East) has proceeded at about the same pace as in Asia, in spite of Africa having a much lower level of industrialization. Asia follows the standard development process: higher income, industrialization, and urbanization all proceed simultaneously. This gives rise to what they refer to as “production cities.” In contrast, in Africa, surplus income from the exports of natural resources leads to greater demand for nontradable goods which are produced in urban centers. This gives rise to what they refer to as “consumption cities.” This disconnect between industrialization and urbanization has also been noted by [Fay and Opal \(2000\)](#) and [Henderson et al. \(2013\)](#).

We now describe the Gollin et al. model in some more detail. They propose a small open economy model with four sectors (food, tradable goods, nontradable goods, and natural resources). By assumption, food production is a rural activity, whereas tradable and nontradable goods are produced in cities. Natural resources are sold internationally and have no domestic market. In this model, a positive shock to natural resources (an increased stock or an increased price) will lead to urbanization without industrialization. Through a standard Rybczynski effect, more labor will be employed in natural resources and less in food and tradable goods. In contrast, because of the positive income effect, the demand for nontradable goods will increase. As a result, the food and tradable good sector will shrink, and the nontradable good sector will expand. If the expansion of the nontradable good sector outweighs the contraction of the tradable good sector, urbanization will increase in the absence of industrialization.

In addition to focusing on the role of natural resources, the work by Gollin et al. also illustrates that comparative advantage and trade can change a country's standard development path. Because of trade, not all countries may need to go through a structural transformation from agriculture to manufacturing as they develop. In that sense, international trade may make the relation between development and the spatial concentration of economic activity more heterogeneous. A broader implication is that using the historical experience of developed countries to "predict" what will happen in developing countries, though useful, should be done with caution. Comparative advantage in early developers may very well be different from comparative advantage in late developers, thus changing the relation between development, industrialization, and urbanization. This connects back to the work of [Matsuyama \(1992\)](#) which we discussed earlier.

That urbanization has proceeded without industrialization does not necessarily imply that urbanization has proceeded without growth. In fact, in Gollin et al., the growth of the urban nontradable good sector is a direct consequence of the positive income shock coming from natural resources. Not everyone agrees though. [Fay and Opal \(2000\)](#), for example, claim that Africa has urbanized in the absence of economic growth. However, given the severe measurement problems that plague income *per capita* in Africa, [Henderson et al. \(2013\)](#) are skeptical of that claim. In fact, when using human capital accumulation, as measured by average number of years in school, they find that the relation with urbanization is not different in Africa compared with the rest of the world.

22.2.2 Development: Continuum of locations

Although increased urbanization is a basic fact of development, limiting the focus to the urban–rural distinction may ignore some of the richer growth dynamics. After all, there are denser and less dense rural areas, and there are bigger and smaller cities. In this section, we take a comprehensive approach. Rather than focusing on cities of different sizes, we focus on all locations. This is important for at least four reasons. First, cities are not islands, and they form part of the overall spatial distribution of the population and economic activity. Second, when going back in time, or when focusing on developing countries, we find the percentage of the population living in rural areas is not trivial. Third, some of the stylized facts that hold for cities may no longer hold when all locations are included. Fourth, when we limit our focus to cities, we introduce a selection bias that we need to be aware of, since by definition cities are locations that benefited from high growth at some point in the past.

In what follows we start by analyzing some of the stylized facts related to growth across locations of different sizes and densities, and then briefly discuss some models that are able to capture the observed dynamics. We also review recent work that focuses on the link between the economy's overall spatial structure and its aggregate growth. It emphasizes the need to develop models that reconcile the main macro growth facts with the observed spatial heterogeneity of economic activity.

22.2.2.1 Facts

22.2.2.1.1 Population growth dynamics and Gibrat's law

Several articles have looked at whether growth is orthogonal to size when considering the entire distribution of locations. [Holmes and Lee \(2010\)](#) divide the United States into a grid of 6 mile by 6 mile squares (93.2 km^2) and find an inverted-U relation between size and the growth of the population between 1990 and 2000. Squares with a population of less than 1000 have an average growth rate of 0.054; growth peaks at 0.149 for squares with a population between 10,000 and 50,000; and growth then declines to 0.06 for squares with more than 500,000 people. This translates into growth peaking in locations with a population density between 100 and 500 per square kilometer. This suggests that Gibrat's law can be rejected when looking at the entire distribution of locations. Using the same time period but focusing on census places, [Eeckhout \(2004\)](#) finds that growth satisfies Gibrat's law. Part of the difference from the findings of [Holmes and Lee \(2010\)](#) is that the census places in [Eeckhout \(2004\)](#) cover only 74% of the US population, leaving out some of the areas with very low population densities.

If there is some doubt about the orthogonality of growth to size in recent times, there is even more doubt when going back in time. Gibrat's law is, if anything, a fairly recent phenomenon. [Michaels et al. \(2012\)](#) use data on US subcounty divisions (in particular, minor civil divisions) to analyze the relation between population density and population growth over the period 1880–2000. As shown in [Figure 22.1](#), the data show a U-shaped relation which becomes flat for high-density locations. Low-density locations, with

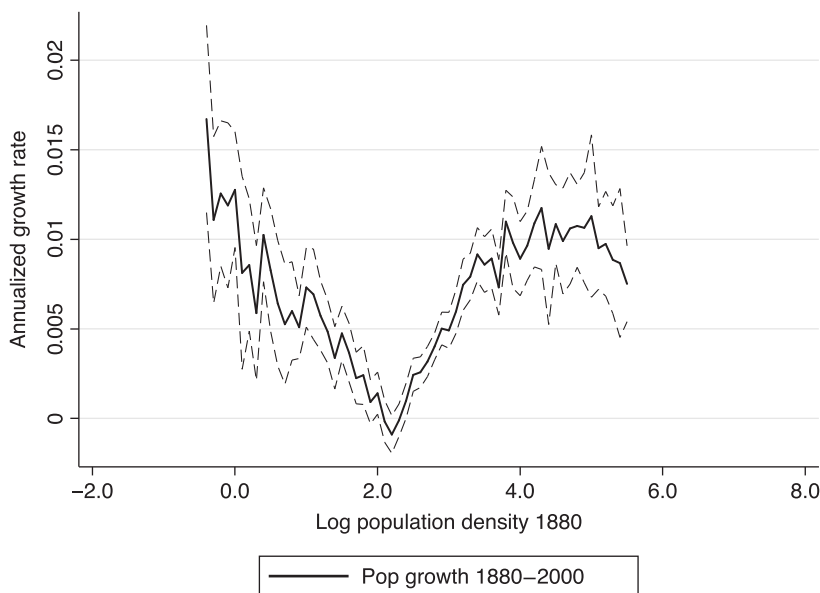


Figure 22.1 Population growth from 1880 to 2000 for US minor civil divisions. Source: [Michaels et al. \(2012\)](#).

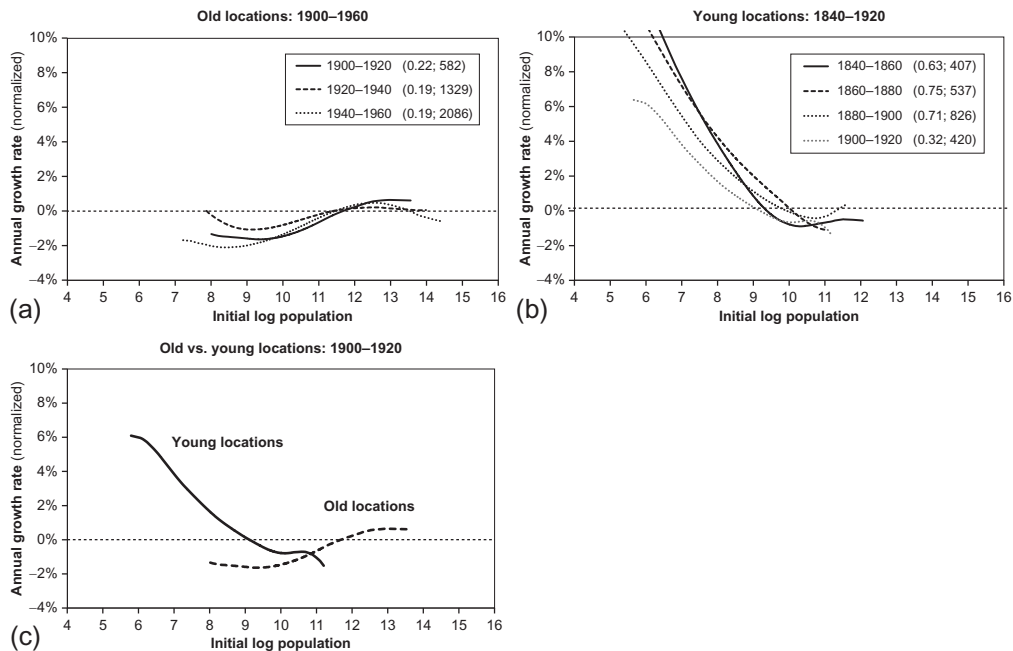
populations of less than 7 per square kilometer in 1880 (less than $\log 2$), exhibit a negative relation between the initial density in 1880 and growth over the period 1880–2000. For medium-density locations, with populations between 7 and 55 per square kilometer (between $\log 2$ and $\log 4$), this relation is positive. It is only for the highest-density locations, with populations above 55 per square kilometer in 1880 (above $\log 4$), that the relation becomes orthogonal. That is, if Gibrat's law holds, it holds only for high-density cities and not for rural areas. This finding illustrates that exclusively focusing on cities is misleading in terms of changes to the whole geography, especially taking into account that about half of the US population in 1880 lived in the intermediate range of locations that experienced divergent growth over the following century. Michaels et al. relate this finding to the structural transformation: divergent growth is most prominent in locations that are transitioning from being agricultural based to being manufacturing based, which reshapes the entire national economic geography.

In a related study, Desmet and Rappaport (2013) use data on US counties from the decennial censuses, starting in 1800, and analyze the relation between size and growth over ten 20-year periods until 2000. They strongly reject orthogonal growth until very recently. Until the 1940s, smaller counties exhibited dispersion (convergence), medium-sized counties exhibited concentration (divergence), and large counties exhibited dispersion (convergence). In more recent time periods, the dispersion at the lower end has disappeared, although the medium-sized counties continue to show some tendency toward further concentration. They show that the nonorthogonality at the lower tail of the distribution is intimately related to the age of a location. Figure 22.2 a and b shows how newly settled (young) locations tend to grow faster than long-settled (old) locations. Young locations exhibit strong convergent growth, whereas old locations exhibit slight divergent growth, except for the largest ones. Although most young locations are also small, not all old locations are large. As can be seen in Figure 22.2 c, the distinction between young and old is therefore not just picking up a size effect.

When the westward settlement of the United States came to an end, convergent growth among smaller locations weakened and disappeared. The importance of settlement for understanding US growth dynamics was emphasized in earlier work by Beeson and DeJong (2002). As for the divergent growth of medium-sized locations, Desmet and Rappaport (2013) relate it to evidence regarding either the declining share of land in production (as in Michaels et al., 2012) or increasing agglomeration economies owing to the introduction of new technologies (as in Desmet and Rossi-Hansberg, 2009).

22.2.2.1.2 Population distribution

Another important finding is that for the last 200 years, the spatial distribution of the population (and population density) has been close to lognormal. This is true, for example, when focusing on the distribution of population levels across US counties as early as 1790, as can be seen in Figure 22.3. The distribution of population densities across minor



A location is "young" if no more than 40 years have passed since the state or territory in which it is located first had two or more counties with positive population. A location is old if more than 60 years have passed since it experienced its final significant geographic change.

Figure 22.2 Population growth from 1800 to 2000 for US counties. Source: *Desmet and Rappaport (2013)*.

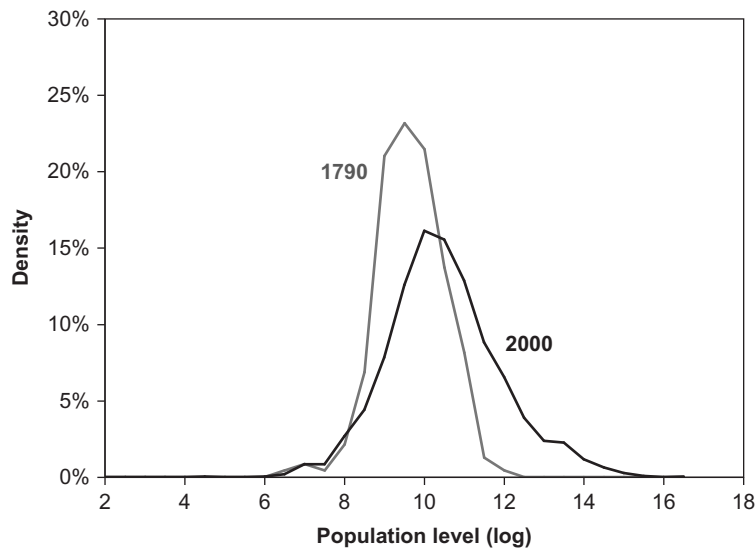


Figure 22.3 Logarithmic population distribution from 1790 to 2000 for US counties. Source: *Desmet and Rappaport (2013)*.

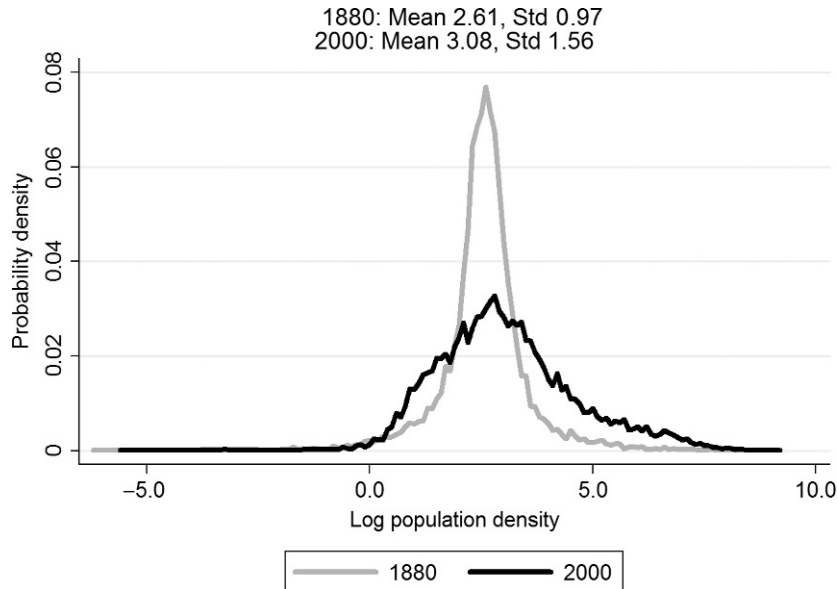


Figure 22.4 Logarithmic population densities from 1880 to 2000 for US minor civil divisions. *Source: Michaels et al. (2012).*

civil divisions in 1880 in [Figure 22.4](#) shows a similar picture. Although the population distribution has essentially remained lognormal (in both levels and densities), the dispersion has increased, mainly because the larger locations today are much larger than they were in the nineteenth century, whereas the smaller locations are not.

22.2.2.1.3 Income growth dynamics

In addition to the focus on population dynamics, some articles have emphasized income *per capita* dynamics. While [Michaels et al. \(2012\)](#) show how the structural transformation can help us understand differential population growth across locations, [Caselli and Coleman \(2001\)](#) argue that the structural transformation can account for the observed income *per capita* convergence across US regions over the last century. Between 1880 and 1980, the South/North relative wage increased from 0.4 to 0.9. In 1880, there was a strong negative correlation between income per worker and the share of employment in agriculture across US states. Over the following century, the states which had most agriculture initially were also the ones where agriculture declined the most and where income *per capita* growth was strongest. These findings taken together, [Caselli and Coleman \(2001\)](#) show that this can explain regional convergence in income *per capita*. This is broadly consistent with evidence provided by [Kim and Margo \(2004\)](#), who show that US income *per capita* diverged across regions during the nineteenth century and early twentieth century, and then started converging dramatically. As in [Caselli and Coleman](#)

(2001), they relate this to changes in industrial structure across regions. During the industrialization of the Northeast and the formation of the manufacturing belt, regional differences in specialization increased, and with it regional differences in income *per capita*. At the beginning of the twentieth century, this trend reversed, and regional specialization started to decline (Kim, 1998).

This pattern of rising regional divergence followed by a process of regional convergence is common across countries. The relation between income *per capita* and regional dispersion in income *per capita* often exhibits an inverted-U-shaped pattern, a phenomenon Kim (2009) refers to as a “spatial Kuznets curve.” In agrarian economies, regional differences are limited. Early industrialization leads to clusters of manufacturing activity emerging in particular locations, leading to an increase in regional income dispersion. As industrialization spreads and agriculture loses importance across the economy, those income differences decline. This pattern has been documented in the 2009 *World Development Report* (World Bank, 2009) for both developing countries and developed countries.

22.2.2.1.4 Relation between spatial agglomeration and growth

The discussion above focused on the relation between development and the convergence (or divergence) of income *per capita* across space. Another, not less important, question is how the overall spatial structure of the economy affects *aggregate*, rather than *local*, growth. Since policy makers often try to affect the spatial distribution of economic activity—as mentioned in Section 22.1, nearly 70% of governments implement policies that slow down urbanization—having a convincing answer to this question would seem to be of much interest. Unfortunately, empirical studies are scarce. One of the few examples is the study of Brühlhart and Sbergami (2009), who use measures of the spatial concentration of employment for a panel of European countries, and find that greater spatial concentration promotes growth up to a GDP *per capita* threshold of around \$12,000 (in 2006 prices). Using urbanization as a proxy for spatial concentration, they find similar results for a large panel of countries across the globe.

22.2.2.2 Theory

22.2.2.2.1 Population growth dynamics and Gibrat's law

There are different possible explanations for why the spatial distribution of economic activity or the population changes with a country's development. Michaels et al. (2012) propose a model that explains how the structural transformation from agriculture to nonagriculture affects the relation between population (or employment) density and growth. Since the timing of the structural transformation is related to an economy's level of development, their theory provides valuable predictions for how spatial growth patterns change along an economy's development path. Their theory also has implications for the evolution of the dispersion of the population over time.

The model consists of a continuum of locations that can produce agricultural or non-agricultural goods using land and labor. Time is discrete and is indexed by t . Workers are perfectly mobile across locations. Workers' preferences between the consumption of agricultural goods, c_A , and the consumption of nonagricultural goods, c_N , are of the constant elasticity of substitution type,

$$(ac_A^\rho + (1-a)c_N^\rho)^{\frac{1}{\rho}},$$

with an elasticity of substitution between both types of goods of less than 1, that $1/(1-\rho) < 1$. The production function is of the Cobb–Douglas type for land and labor. Output in sector j and sector i in period t is given by

$$Y_{jit} = L_{jit}^{\eta_j} \Gamma_{jt} \theta_{jit} L_{jit}^{\mu_j} H_{jit}^{1-\mu_j},$$

where L_{jit} and H_{jit} denote land and labor used, and where TFP depends on external economies of scale, $L_{jit}^{\eta_j}$, a sectoral productivity component common across locations, Γ_{jt} , and a location-specific sectoral productivity component, θ_{jit} . Agriculture is assumed to be more land intensive than nonagriculture, so $\mu_A < \mu_N$, and agriculture benefits less from agglomeration economies than nonagriculture, so $\eta_A < \eta_N$. The location-specific sectoral productivity component, θ_{jit} , is hit by idiosyncratic shocks ϕ_{jit} :

$$\theta_{jit} = \phi_{jit} \theta_{jit-1}^{\nu_j},$$

where the parameter ν_j is key, as it is inversely related to the mean reversion in location-specific productivity. In particular, if $\nu_j = 0$, there is no persistence in productivity, so we get high mean reversion; if $\nu_j = 1$, there is high persistence in productivity, so we get low mean reversion. It is assumed that mean reversion in agriculture is greater than in non-agriculture, so $\nu_A < \nu_N$.

Since workers can costlessly relocate, utility equalizes across locations, and the dynamic equilibrium collapses to a sequence of static equilibria. As long as agglomeration economies are not too strong compared with dispersion forces coming from land, the equilibrium of the economy is stable and unique. The theory generates the following results. First, population density is lower in locations specialized in agriculture than in those specialized in nonagriculture. This reflects the relatively higher land intensity in agriculture ($\mu_A < \mu_N$) and the relatively weaker agglomeration forces in that same sector ($\eta_A < \eta_N$). Second, the dispersion in population density is greater across nonagricultural locations than across agricultural locations. This is a consequence of the weaker mean reversion in nonagriculture, which implies the variance and the maximum value of productivity is greater in nonagriculture than in agriculture. Third, the structural transformation displaces the population from agricultural locations to nonagricultural locations, and also makes some locations switch from agriculture to nonagriculture. Relative increases in agricultural productivity, together with an elasticity of substitution of less than 1 between agricultural and nonagricultural goods, underlie this phenomenon.

These predictions are able to explain two of the more relevant features of the data. One is the increased dispersion in population density over time, as shown in [Figure 22.4](#). The greater relative dispersion in nonagricultural employment density implies that the structural transformation away from agriculture increases the overall dispersion in population density. Another is the nonlinear relation between the initial population density and growth: as shown in [Figure 22.1](#), for low-density locations the relation is negative, for medium-density locations the relation is positive, and for high-density locations the relation is orthogonal. On the one hand, for agricultural locations (which tend to be low-density places), strong mean reversion in productivity implies a negative relation between density and growth. Among those locations, the higher-density ones are those which had the highest productivity shocks in the past; in the presence of mean reversion, their relative productivity (and density) is therefore likely to go down. On the other hand, for nonagricultural locations (which tend to be high-density places), there is no relation between density and growth. The absence of mean reversion as $\nu_N \rightarrow 1$ implies that growth is essentially orthogonal to density, so Gibrat's law holds for those locations. In between those two extremes, we have the nonspecialized medium-density locations where the share of agriculture, on average, decreases with the initial density. The structural transformation leads to greater population growth in those locations with a higher proportion of nonagriculture, thus implying a positive relation between the initial population density and growth.

An alternative explanation by [Desmet and Rappaport \(2013\)](#) focuses on transition dynamics and entry. In their one-sector model, locations gradually enter over time. Upon entry, they draw a productivity from a distribution. Frictions on positive population growth slow the upward transition to each location's steady state and so cause population growth from low levels to be characterized by convergence. The congestion arising from the fixed supply of land in each location gradually diminishes over time. This is consistent with either a decrease in land's share of factor income (as in [Michaels et al., 2012](#)) or an increase in the effect of agglomeration on productivity (as in [Desmet and Rossi-Hansberg, 2009](#)). As this allows steady-state population levels to become more sensitive to underlying differences in exogenous productivity, it introduces a force toward divergence. Once entry is complete and the degree of net congestion has stabilized, the assumed orthogonality of productivity growth causes population growth to be orthogonal as well.

22.2.2.2.2 Income growth dynamics

The models mentioned above remain silent on income *per capita* differences across space, essentially because there is only one type of labor and all workers are perfectly mobile across locations. [Caselli and Coleman \(2001\)](#) introduce different skill types. Although workers are geographically mobile, regional differences in skill composition will lead to income *per capita* differences.

To be more precise, they propose a North–South model of the structural transformation with three basic assumptions. First, TFP growth is higher in agriculture than in manufacturing. The production technologies in food and manufacturing in region i and time t use land (T), labor (L), and capital (K) and are of Cobb–Douglas type of the form

$$F_t^i = A_{ft}^i (T_{ft}^i)^{\alpha_T} (L_{ft}^i)^{\alpha_L} (K_{ft}^i)^{1-\alpha_T-\alpha_L}$$

and

$$M_t^i = A_{mt}^i (T_{mt}^i)^{\beta_T} (L_{mt}^i)^{\beta_L} (K_{mt}^i)^{1-\beta_T-\beta_L},$$

where the South has a comparative advantage in agriculture and the North has a comparative advantage in manufacturing. As mentioned before, it is assumed that (exogenous) TFP growth in agriculture, g_f , outpaces that in manufacturing, g_m .

Second, there is a cost of acquiring nonfarm skills, and this cost drops over time. The demographic structure is that of a dynasty, with a constant population and a probability of death in each period. In each period each person is endowed with one unit of time. When born, a person decides whether to immediately start working on the farm, or to first spend ξ, ζ^i units of time getting trained to work in manufacturing, where ξ_t captures the economy's overall efficiency in providing training and ζ^i is distributed among the people of a generation according to a time-invariant density function $\mu(\zeta^i)$. Assuming that ξ_t drops over time implies that training becomes cheaper over time. As a result, the cutoff ζ^i below which individuals invest in skill acquisition rises over time, implying more people become skilled.

Third, the income elasticity of demand for agricultural goods is less than 1. In particular, the period utility derived from consuming food, c_f , and manufactured goods, c_m , is

$$u(c_{ft}^i, c_{mt}^i) = \frac{((c_{ft}^i - \gamma)^\tau (c_{mt}^i)^{1-\tau})^{1-\sigma}}{1-\sigma},$$

where $\gamma > 0$ is the subsistence constraint on food consumption, implying the less than unit income elasticity of demand for food.

Because of the initially high cost of acquiring nonfarm skills, the relative supply of manufacturing workers is low, implying a substantially higher manufacturing wage. Given that the South has a comparative advantage in agriculture, this implies a wage gap in favor of the North, in spite of labor being mobile across regions. As the overall economy becomes richer because of general productivity growth, the demand for manufacturing goods increases, shifting labor from agriculture to manufacturing. This process is further reinforced by the faster TFP growth in agriculture compared with manufacturing. With a declining weight of agriculture in the economy, average wage differences across regions drop. The falling cost of acquiring nonfarm skills enhances this convergence across regions and has the additional advantage of leading to a reduction in

wage differences not just across regions, but also within regions between farm and non-farm workers. It is this latter feature which the model would not be able to capture if it did not assume a falling cost of acquiring manufacturing skills.

Whereas this model predicts that the structural transformation leads to income convergence across regions, it is likely that in the early stages of industrialization the opposite happened. In the model, part of the convergence between North and South happens because average wages converge as a result of the sectoral composition becoming more similar across regions. During the early stages of industrialization, when the North shifted increasingly into manufacturing, the opposite should have happened. As mentioned before, this would be consistent with the evidence in [Kim and Margo \(2004\)](#), who describe a process of income divergence during the nineteenth century, followed by convergence, which is particularly strong during the second half of the twentieth century.

An assumption in most of these models is that labor is freely mobile across regions. This does not necessarily contradict the evidence of nominal and real wages being substantially higher in the West than in the rest of the country during the nineteenth century ([Easterlin, 1960](#); [Rosenbloom, 1990](#); [Mitchener and McLean, 1999](#)). As in [Caselli and Coleman \(2001\)](#), this gap might be due to differences in skills. This does not seem to be the entire story though, since these differences also existed within occupations. Focusing on 23 occupations, [Rosenbloom \(1990\)](#) documents within-occupation average real wage differences of more than 50% between the West and the South in 1870; by the end of the nineteenth century, this difference continued to exist, although it had been cut in half. This suggests that labor markets were not completely integrated, and that moving costs were driving a wedge between wages in the West and the rest of the country.

22.2.2.2.3 Gibrat's law and Zipf's law

An interesting related question is how Gibrat's law is connected to Zipf's law. Theory says that proportionate (or random) growth should give rise to a lognormal distribution ([Gibrat, 1931](#)). That is, Gibrat's law implies a lognormal distribution. Consistent with this, [Eeckhout \(2004\)](#), using data on census places, shows that growth between 1990 and 2000 satisfies Gibrat's law and that the size distribution of places is lognormal. Since the lognormal distribution and the Pareto distribution are very different, Gibrat's law seems to be inconsistent with the observation that the city-size distribution conforms to Zipf's law.¹ The puzzle is partly resolved when it is realized that cities make up the upper tail of the size distribution of all locations, and at that upper tail the lognormal distribution is actually very similar to the Pareto distribution. So although Gibrat's law does not imply a Pareto distribution overall, in the upper tail they are similar (see [Ioannides and Skouras, 2013](#) for a further discussion). Note that there are restrictions on the stochastic process which can lead Gibrat's law to imply Zipf's law. For example, [Gabaix \(1999\)](#) shows that if cities cannot fall below a minimum size, then Gibrat's law implies a city

¹ [Section 22.4.2.1](#) has a longer discussion on city-size distributions.

size distribution that converges to Zipf's law. The intuition is simple: we get the density function peaking at the minimum city size and at the same time the lower bound on size pushes more cities to become large, implying the fatter upper tail, characteristic of Zipf's law (see [Duranton and Puga, 2014](#) for a review of this literature).

Returning to the observed lognormality in the size distribution of all places, an open question is whether the lognormality is due to past proportionate (or random) growth or whether it is due to some underlying lognormal distribution of locational characteristics. The finding in both [Michaels et al. \(2012\)](#) and [Desmet and Rappaport \(2013\)](#) that the orthogonality of growth across locations in the United States was categorically violated until recently sheds doubt on whether random growth can have caused the present-day lognormal distribution of the population. This doubt is further enhanced once we observe that the distribution of the population in 1790 was already lognormal, as shown in [Figure 22.3](#). More consistent with the observed growth rates is that the combined underlying determinants of the steady-state population are distributed lognormally ([Krugman, 1996](#); [Rappaport and Sachs, 2003](#)). This does not require any one characteristic of a location to be distributed lognormally. As shown by [Lee and Li \(2013\)](#), as long as there are enough factors, the population distribution will be lognormal, even if none of the factors individually is lognormally distributed.

22.2.2.2.4 Spatial agglomeration and aggregate growth

So far we have analyzed spatial growth patterns, but we have not focused on the relation between space and aggregate growth. That is, how does the overall spatial structure of an economy affect its aggregate growth rate? There exist some dynamic extensions of two-region *new economic geography* models which were reviewed in the previous edition of this handbook ([Baldwin and Martin, 2004](#)). Although these extensions analyze the relation between geography and growth, their focus on a small number of locations limits their ability to capture the overall spatial distribution of the economy.

Incorporating a richer spatial structure into dynamic growth models is complex because it increases the dimensionality of the problem. As discussed in [Desmet and Rossi-Hansberg \(2010\)](#), models become quickly intractable and unsolvable when agents' decisions depend on the distribution of economic activity over both time and space. In recent years, some attempts have been made to incorporate forward-looking agents into models with a continuum of locations ([Brock and Xepapadeas, 2008](#); [Boucekkine et al., 2009](#); [Brock and Xepapadeas, 2010](#)). Unfortunately, to keep these spatial dynamic models solvable, they do not take into account many relevant spatial interactions, such as transportation costs and factor mobility.

To get around this problem, [Desmet and Rossi-Hansberg \(2014a\)](#) impose enough structure so that future allocation paths do not affect today's decisions. This result is obtained by assuming that workers are freely mobile and that innovation by firms diffuses across space. The model strikes a balance between being tractable and having a rich spatial structure that allows it to connect with the data. They use their theoretical framework to

study the spatial and aggregate evolution of the US economy over the last half century. To highlight some of the main features of the model, we present here a simplified one-sector version. Land and agents are located on the unit interval $[0, 1]$, time is discrete, and the total population is $\bar{L}$. Agents solve

$$\begin{aligned} & \max_{\{c(\ell, t)\}_0^\infty} E \sum_{t=0}^{\infty} \beta^t U(c(\ell, t)) \\ & \text{subject to} \\ & w(\ell, t) + \frac{\bar{R}(t)}{\bar{L}} = p(\ell, t)c(\ell, t), \quad \text{for all } t \text{ and } \ell, \end{aligned}$$

where $c(\ell, t)$ is consumption at location ℓ and time t , $p(\ell, t)$ is the price of the consumption good, $w(\ell, t)$ denotes the wage, and $\bar{R}(t)$ are total land rents, so $\bar{R}(t)/\bar{L}$ is the dividend from land ownership, assuming that agents hold a diversified portfolio of land. Free mobility implies that utility equalizes across locations.

Firms use land and labor to produce. Production per unit of land at location ℓ at time t is $Z(\ell, t)L(\ell, t)^\mu$, where $\mu < 1$, $Z(\ell, t)$ denotes TFP, and $L(\ell, t)$ is the amount of labor per unit of land used. A firm's TFP depends both on technology diffusion and on innovation decisions. Technology diffuses between time periods. Before the innovation decision at time t , a firm at location ℓ has access to

$$Z^-(\ell, t) = \max_{r \in [0, 1]} e^{-\delta|\ell-r|} Z^+(r, t-1), \quad (22.3)$$

where the “minus” superscript in Z^- refers to the technology a location has access to before innovation, whereas the “plus” superscript in Z^+ refers to the technology a location ends up using after the innovation decision. In addition to the technology it gets access to through diffusion, a firm can decide to buy a probability $\phi \leq 1$ of innovating at cost $\psi(\phi)$. A firm that obtains the chance to innovate draws a technology multiplier z from a Pareto distribution with shape parameter a and lower bound 1, so the expected technology for a given ϕ is

$$E(Z^+(\ell, t) | Z^-) = \left(\frac{\phi + a - 1}{a - 1} \right) Z^-.$$

The innovation draws are independent and identically distributed across time, but not across space. Hence, innovation draws are spatially correlated, and firms that are located arbitrarily close to each other obtain exactly the same innovations. The timing of the problem is as follows. During the night, between periods $t-1$ and t , technology diffuses locally. This leads to a level of technology $Z^-(\ell, t)$ in the morning. Each firm then decides on how many workers it wants to hire, how much it wants to bid for land, and how much to invest in innovation. Only the firm that offers the highest bid for land in a given location gets to rent the land. Investment in innovation, if it occurs, then leads to a new technology, $Z^+(\ell, t)$. Production happens at the end of the period.

We now turn to the firm's problem. The objective function of a firm in a given location ℓ at time t_0 is

$$\max_{\{\phi(\ell, t), L(\ell, t)\}_{t_0}^{\infty}} E_{t_0} \left[\sum_{t=t_0}^{\infty} \beta^{t-t_0} \left(p(\ell, t) \left(\frac{\phi(\ell, t)}{a-1} + 1 \right) Z^-(\ell, t) L(\ell, t)^{\mu} - w(\ell, t) L(\ell, t) - R(\ell, t) - \psi(\phi(\ell, t)) \right) \right],$$

where β is the discount factor and $R(\ell, t)$ is the firm's bid rent, which is chosen to maximize the probability of winning the auction to rent land. As discussed in [Desmet and Rossi-Hansberg \(2012\)](#), in this setup firms invest in innovation, in spite of operating in a perfectly competitive market, because it allows them to bid a higher price for land. Returning to the above maximization problem, we recall that labor is freely mobile and that firms compete for land and labor every period with potential entrants that, because of diffusion, have access to the same technology. The decision on how many workers to hire and how much to bid for land are therefore static problems. The only problem that is in principle dynamic is the innovation decision, but here as well the dynamic problem simplifies to a static one. The continuity in the diffusion process and the spatial correlation in innovation realizations guarantee that a firm's decisions do not affect the expected technology it wakes up with tomorrow. Hence, future allocation paths do not affect a firm's decision today. This key result is what makes the dynamic spatial model solvable and computable.

The importance of this framework is that it not only has implications for the interaction between density and growth at the local level, but it also analyzes the interaction between the spatial distribution of economic activity and aggregate growth. When applying their framework to the evolution of the US economy in the last 50 years, [Desmet and Rossi-Hansberg \(2014a\)](#) can account for the main spatial patterns, such as the evolution in the dispersion of land prices, as well as the main macroeconomic stylized facts, such as the evolution of aggregate growth and wages. More broadly, the aim is to develop a unified framework to study the interaction between space and the macroeconomy. In other work, the same authors use a similar setup to quantitatively analyze the impact of global warming on both the spatial distribution of economic activity and global welfare ([Desmet and Rossi-Hansberg, 2014b](#)).

22.3. DEVELOPMENT, SPACE, AND INDUSTRIES

Although we touched upon the structural transformation from agriculture to manufacturing, our main focus in the previous section was on aggregate population growth across different locations. In this section, we delve deeper into the incentives of different industries to concentrate or disperse, and analyze the geography of sectoral employment growth. In particular, we are interested in the differences between manufacturing and services.

This is related to the broader question of how the spatial distribution of economic activity changes with development, for at least two reasons. First, if spatial growth patterns differ across sectors, then a country's overall spatial organization will change as it develops and the relative importance of different sectors changes. Second, for a given sector, spatial growth patterns may also change over time, as sectors transition from being young to being maturer. In what follows, we discuss some of the recent empirical findings, as well as theories that can account for them.

22.3.1 Manufacturing versus services

In recent decades, US manufacturing has become spatially more dispersed and services have become spatially more concentrated. On the basis of US county employment data between 1970 and 2000, [Table 22.1](#) shows that the difference in the logarithm of employment between the 70th percentile and the 30th percentile decreased in manufacturing and increased in services. This implies manufacturing became more equally spread across US counties, whereas the opposite happened to services. When the standard deviation of the logarithm of employment is used as an alternative measure of the degree of concentration, the result is similar. Since services started off being less concentrated than manufacturing, this implies services becoming more like manufacturing in their degree of spatial concentration.

Does this mean that manufacturing is dispersing and services are concentrating across all locations? To get a more precise idea, [Desmet and Fafchamps \(2006\)](#) and [Desmet and Rossi-Hansberg \(2009\)](#) run nonlinear kernel regressions of the form

$$L_{t+s}^i = \phi(L_t^i) + e_t^i,$$

where L_t^i is the logarithm of employment in year t and county i . [Figure 22.5](#) shows that the tendency toward greater geographic dispersion in manufacturing is happening across the entire distribution. Counties with less manufacturing have been experiencing faster manufacturing employment growth than those with more manufacturing. In the case of services, the picture is more complex: the relation between size and growth is S shaped.

Table 22.1 Spatial concentration of employment (as a logarithm)

	1970	2000
Difference between 70th percentile and 30th percentile		
Manufacturing	1.81	1.74
Services	1.29	1.52
Standard deviation		
Manufacturing	2.05	1.89
Services	1.40	1.52

Source: REIS, Bureau of Economic Analysis.

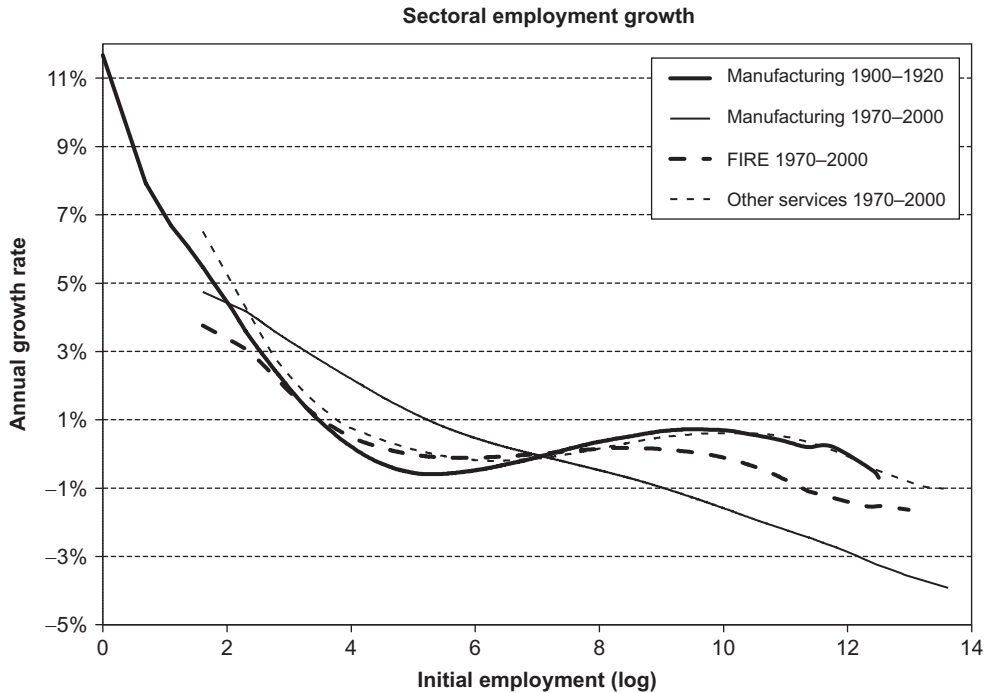


Figure 22.5 Sectoral employment growth in US counties. Source: *Desmet and Rossi-Hansberg (2009)*.

The tendency toward a greater geographic concentration of services employment has mainly affected mid-sized service employment clusters. European regions look similar to US counties: deconcentration in manufacturing and greater concentration in services.

22.3.2 Life cycle of industries and spatial distribution

One possible explanation for this trend is the greater land intensity of services. As services compete for the same land as manufacturing in high-density urban environments, services are increasingly competing manufacturing out of cities. Another possible explanation has to do with the life cycle of an industry. Younger industries stand more to gain from knowledge spillovers, which are enhanced by the geographic concentration of economic activity. This could help us understand the recent tendency toward greater concentration in services. As shown by *Hobijn and Jovanovic (2001)*, the impact of information technology (IT) is greater in services than in manufacturing. They compute IT intensity—the share of IT equipment in the total stock of equipment—in different sectors in 1996, and find a figure of 42.4% in services and a much lower 17.9% in manufacturing. Using alternative definitions of the importance of IT, *Triplett and Bosworth (2002)* and *Basu and Fernald (2007)* find similar differences between manufacturing and services.

To operationalize the idea of the age of an industry, [Desmet and Rossi-Hansberg \(2009\)](#) propose using the time elapsed since the introduction of a general-purpose technology (GPT). [David and Wright \(2003\)](#) and [Jovanovic and Rousseau \(2005\)](#) argue that the two major GPTs of the twentieth century were electricity and IT. As for their timing, [Jovanovic and Rousseau \(2005\)](#) propose identifying the starting date of a GPT by taking the year in which it reaches a 1% diffusion. In electricity, this corresponds to 1894, the year of the first hydroelectric facility at Niagara Falls, and in IT this corresponds to 1971, the year of the Intel 4004 microprocessor. As the ending date of a GPT, they take the year when the diffusion curve becomes more or less flat. In the case of electricity, this corresponds to 1929, whereas in IT that point has not been reached yet. Roughly speaking, this makes the period between 1900 and 1920 for electricity comparable to the period between 1970 and 2000 for IT.

While IT is viewed as mainly affecting services, electricity's impact was mostly felt in the manufacturing sector ([David and Wright, 2003](#)). If age plays an important role in the spatial growth patterns of an industry, we would expect the spatial growth pattern of manufacturing at the beginning of the twentieth century to look similar to that of services at the end of the twentieth century. As seen in [Figure 22.5](#), this is indeed the case. The spatial growth pattern of manufacturing at the beginning of the twentieth century looks very different from that of manufacturing at the end of the twentieth century, but very similar to that of services at the end of the twentieth century. This finding implies that when analyzing the relation between space and growth, not only the sectoral composition of the economy matters but also the age of the different sectors plays a role. There is nothing inherent about manufacturing exhibiting a tendency toward greater dispersion; indeed, when the sector was young, it became increasingly concentrated.

Motivated by this evidence, [Desmet and Rossi-Hansberg \(2009\)](#) provide a theory for how an industry's spatial growth is related to its life cycle. The model has three forces. First, local knowledge spillovers constitute an agglomeration force that incentivizes the geographic concentration of economic activity. Second, crowding costs coming from land constitute a dispersion force. Third, technology diffuses over space. This constitutes an additional dispersion force. The relative importance of these three forces will depend both on a location's size and on an industry's age. Together, they will be able to capture how the scale dependence of an industry's growth evolves over its life cycle.

To see this, it will be convenient to distinguish between three types of locations in terms of their employment size: small locations, medium-sized locations, and large locations. In small locations, knowledge spillovers have little bite, so TFP is determined by the technology a location has access to through diffusion. Hence, among smaller places, we should see increasing divergence. In medium-sized locations, knowledge spillovers become the dominating force. With congestion forces still being weak, we see increasing concentration among medium-sized places. In large locations, local congestion forces start to dominate knowledge spillovers from neighboring locations. Among those large places, we should thus see increasing dispersion.

The above description suggests an S-shaped relation between size and growth: dispersion among both small and large locations, and concentration among medium-sized locations. Going back to [Figure 22.5](#), we see this description is consistent with the growth patterns of services in recent decades and manufacturing at the beginning of the twentieth century. We referred to those as “young” industries at the beginning of their life cycle. In contrast, “maturer” industries, such as manufacturing in recent decades, exhibit increased dispersion across all sizes. The absence of increased concentration in medium-sized locations reflects knowledge spillovers being less likely to outdo the productivity a location has access to through diffusion. Medium-sized locations that could benefit from knowledge spillover from neighboring locations have already done so, and no longer benefit from further increases in their productivity. As this happens, the upward-sloping part of the S-shaped relation between size and growth, present in younger industries, disappears as industries reach the later stages of their life cycle.

Of course, services and manufacturing are broad sectors; not all subsectors of services benefit from IT in the same way. With a focus on two-digit Standard Industrial Classification (SIC) subsectors of services, evidence obtained by [Chun et al. \(2005\)](#), [Caselli and Paternò \(2001\)](#), and [McGuckin and Stiroh \(2002\)](#) suggests that the most IT-intensive sector is legal services and the least IT-intensive sector is auto repair. Using employment at the two-digit SIC level from the County Business Patterns dataset spanning the period 1977–1997, [Figure 22.6](#) shows employment growth in legal services and auto repair. As expected, legal services exhibit the S-shaped spatial growth pattern. In contrast, auto repair looks like a mature sector, with convergence across the entire distribution. In the same way that not all service activity is concentrating, not all manufacturing is dispersing. We would expect manufacturing activities that most strongly benefit from knowledge spillovers to have less of an incentive to disperse. This explains the findings of [Fallah and Partridge \(2012\)](#), who show that high-tech manufacturing pays a relatively higher price for remoteness. In particular, a 1 km increase in the distance from the nearest metropolitan area decreases employment growth by 0.2% in high-tech manufacturing, compared with a 0.1% decrease in manufacturing overall. We would therefore expect high-tech manufacturing sectors to remain more clustered than the rest of the industry.

The more general link between an industry’s life cycle and its spatial distribution has also been analyzed by [Henderson \(2010\)](#), who provides evidence of standardized manufacturing dispersing and high-tech manufacturing concentrating. In the specific case of the Internet, [Forman et al. \(2005\)](#) show that its use diffused rapidly across the United States, but its more complex applications, such as e-commerce, predominantly located and developed in cities, where there was more easy access to complementary inventions and activities.

The pattern of spatial concentration followed by spatial dispersion as industries mature has been noted in other countries. For example, in [Section 22.4](#), we will discuss data which indicate that Seoul transformed from being a manufacturing center to a service center from 1970 on. Similarly to what happened in the United States and Europe,

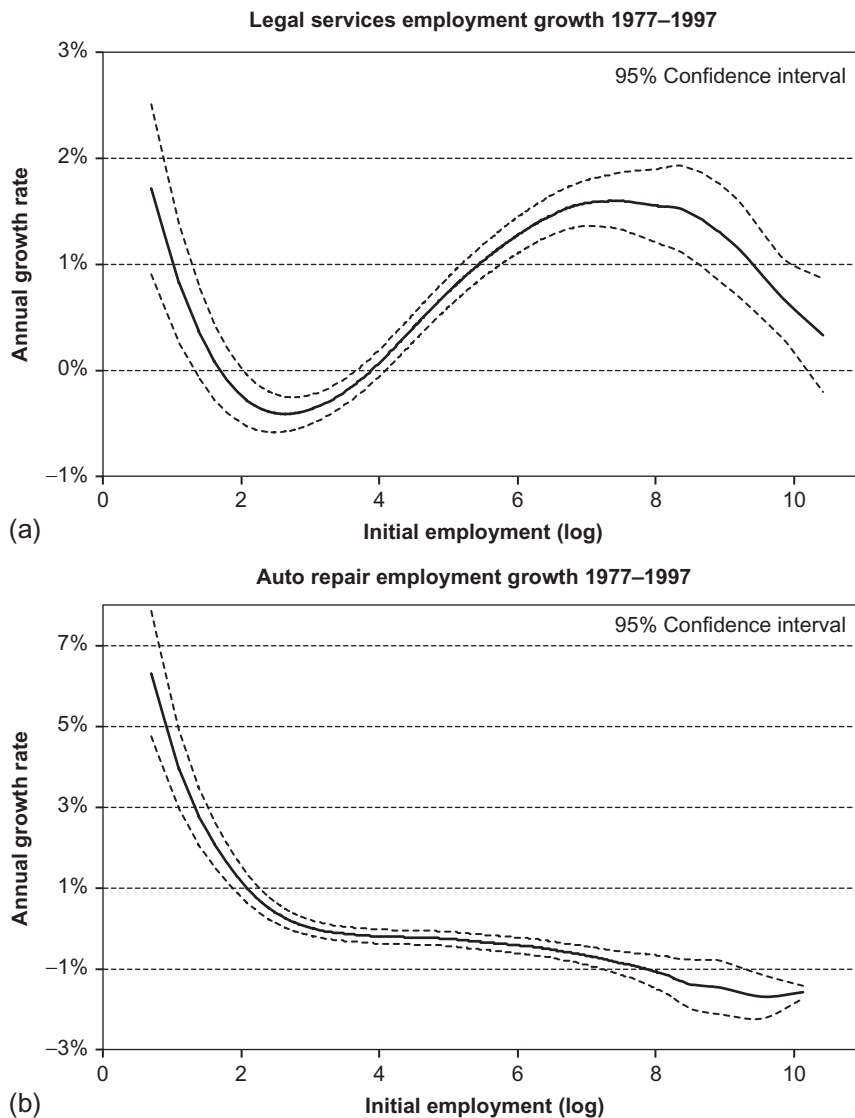


Figure 22.6 Sectoral employment growth in US counties: (a) from Desmet and Rossi-Hansberg (2009) and (b) from Desmet and Rossi-Hansberg (2009).

the loss of manufacturing employment in Seoul benefited the rural areas and the small towns, which experienced an industrial transformation after 1980. Similarly, in the 1990s, the correlation between the manufacturing-service ratio and the size of a city in China was -0.20 , implying that larger cities were relatively more service oriented (Au and Henderson, 2006a). Consistent with this, China's 2008 economic census indicates that telecommunications, software, information, and broadcasting services are

highly concentrated at the upper end of the size distribution of counties. It is, of course, important to note that the timing of these transformations may differ across countries. For example, whereas in the United States manufacturing had become a mature industry by, say, the 1960s, in South Korea this same stage was reached only in, say, the 1980s. This underscores a point we made before: to understand the relation between development and space, it is important to know not just the relative sizes of different sectors but also their ages.

The appearance of clusters during the early stages of an industry's life cycle is not a recent phenomenon. [Trew \(2014\)](#), for example, documents the emergence of industrial hot spots in nineteenth century England. In 1750, two counties in England, Lancashire and the West Riding, had between 65% and 70% of all employment in the country's secondary sector. These were not necessarily the densest areas initially, but they experienced tremendous population growth as the industrial revolution took off. In the nineteenth century, Sheffield, for example, grew from a town of 60,000 inhabitants to a large city of 450,000 inhabitants. London, the country's biggest city, was also a major manufacturing center, as were some of the other large cities, such as Manchester and Birmingham ([Shaw-Taylor and Wrigley, 2008](#)).

22.3.3 Ruralization versus suburbanization

Although manufacturing clusters are spreading out, they often do not move far away. If so, manufacturing growth should be lower in the clusters themselves but higher in areas close to the clusters. Using data on US counties for the last three decades of the twentieth century, [Desmet and Fafchamps \(2005\)](#) find exactly this. In particular, having 1% more manufacturing employment locally lowered manufacturing employment growth by around 2% annually, whereas having 1% more manufacturing employment 40–50 km away increased manufacturing employment growth by 0.1–0.2% annually. These figures refer to manufacturing clusters, rather than to aggregate clusters.

When we look at total employment, the tendency of manufacturing is to suburbanize rather than to ruralize. If so, manufacturing growth should be relatively low in locations with high aggregate employment and relatively high in locations close to aggregate clusters. Again, this is what [Desmet and Fafchamps \(2005\)](#) find. Having 1% more total employment locally lowered manufacturing employment growth by around 0.2% annually, whereas having 1% more manufacturing employment 40–50 km away increased manufacturing employment growth by a little less than 0.01% annually. Though small, the effects are statistically significant, and amount to something much larger once we take into account that we are looking at average annual growth over a period of three decades.

22.3.4 The cost of remoteness

The general tendency toward greater dispersion is mitigated in several ways. First, as already mentioned, high-tech manufacturing tends to remain clustered in high-density

areas to take advantage of knowledge spillovers. Second, the cost of remoteness differs not only across sectors (high-tech vs. low-tech sectors) but also across functions within sectors. With the fragmentation of the value chain, we are witnessing firms locating headquarters and business services in larger cities and production facilities in smaller cities. The evidence for this is reviewed in the next section when models of functional (as opposed to product) specialization by cities are discussed. In general, since 1950, larger cities have moved toward management activities as opposed to production activities, while smaller cities have moved in the opposite direction (Duranton and Puga, 2005). Although the fragmentation of the value chain and the spatial division of labor respond to standard forces of comparative advantage, there are limits to their scope. For example, Tecu (2013) finds that an average US chemical firm is 1.8% more productive in R&D (in terms of patents) if it increases the number of production workers by 10% in the same metropolitan statistical area. In the average metropolitan statistical area, having an average-sized production facility increases the productivity of R&D by 2.5 times in the chemical industry. Doubling the number of production workers has nearly as large an effect on a firm's R&D productivity as doubling the total number of patents generated in the metropolitan statistical area.

The trade-off between moving to cheaper locations and benefiting from proximity may explain the tendency of the different units of multiestablishment firms to locate not too far from each other. In the UK manufacturing industry, for example, establishments that belong to the same firm tend to cluster no more than 50 km from each other, whereas there is no evidence of such clustering by establishments that are part of different firms (Duranton and Overman, 2008).

22.4. THE URBAN SECTOR

Sections 22.2 and 22.3 started with the urban–rural divide and then turned to an analysis of the evolution of economic activity across the continuum of space in a country, moving from the least to the most densely populated locations. For the continuum, the focus was on the spatial transformation in uses of these spaces: how the distribution of the population and the distribution of industrial and service activities change across the continuum with economic growth and technological change. This section has a narrower focus, which is the subject of a large body of literature. We look at the right tail of the continuum in the denser locations that are typically labeled as urban. Because of the sheer volume of the population living at high densities in this tail, it is often the focus of public policy and institutional reform initiatives, as well as people's images of other countries as defined by their largest cities.

This right tail, the urban sector, consists of a hierarchy of cities of very different sizes and functions that transform over time, as suggested before by the results for the continuum. Within the urban sector, cities specialize relatively, and to some degree absolutely,

in particular export activities, giving cities different sizes and different compositions of production activities, occupations, and functions. There are strong patterns in the variation of compositional specialization across the urban hierarchy by city size at a point in time, as well as variation over time within the urban hierarchy depending on the level of economic development. This urban literature has traditionally focused both on the reasons for and the extent of extreme agglomeration and on analyzing why production activities and occupations vary across the hierarchy and over time. There are a number of chapters in prior handbooks which detail work in the literature up to the early years of the twenty-first century (e.g., [Abdel-Rahman and Anas, 2004](#); [Duranton and Puga, 2004](#); [Fujita et al., 2004](#); [Gabaix and Ioannides, 2004](#); [Holmes and Stevens, 2004](#); [Henderson, 2005](#)). We focus on developments since then.

In [Section 22.4.1](#), we start by reviewing some basic facts on specialization within urban hierarchies in different countries today, and then turn to a discussion of models that capture key aspects of the industry, occupation, and functional specialization we see across parts of the urban hierarchy. In [Section 22.4.2](#), we take a more dynamic look, building on the analysis of the structural transformation in [Section 22.3](#). We look at how the products and functions of bigger versus smaller cities have altered dramatically over the last 25 years in particular countries, both developed and developing, with aspects of that transformation depending on the stage of economic development. In [Section 22.4.3](#), we turn to an examination of some policies which have strong effects on aspects of a country's urban hierarchy and thus may affect the relative efficiency of the spatial organization of production.

22.4.1 Production patterns in the urban hierarchy

22.4.1.1 Facts

Older work characterized product specialization in two ways. One way was by using cluster analysis to classify cities as steel cities, auto cities, wood product cities, and the like. The second way was to see how the elasticities of sectoral employment with respect to city size differ across sectors ([Henderson, 1997](#); [Kolko, 1999](#); [Black and Henderson, 2003](#)). For the United States some facts emerge. Small and medium-sized cities were historically relatively specialized in manufacturing, but that degree of specialization has declined as the country has deindustrialized. Specialization in standardized services by smaller and medium-sized cities has increased. Bigger cities have a much more diverse industrial base, and they are much more engaged in the provision of more sophisticated business and financial services.

Here we evaluate more recent work. For developed countries, recent contributions characterize specialization not just by products but also by functions and occupations, with the idea that bigger cities are specialized more in more highly skilled occupations and functions. Although specialization may still be characterized by elasticities of sectoral employment with respect to city size to show what is produced more in different parts of the urban hierarchy, the literature now also uses spatial Gini or Krugman indices to

characterize the degree to which different cities are specialized (without reference to what they are specialized in *per se*). Another body of literature, which we do not review here, focuses on identifying which specific industries are more clustered in space versus more diffusely spread out, compared with a random allocation across space (Ellison and Glaeser, 1997; Duranton and Overman, 2005).

For developing countries, there are a few recent articles looking at specialization that offer a somewhat different perspective. One more innovative article focuses on a different dimension: the division of labor between and within cities, as it varies across a less developed hierarchy. In Section 22.4.2, we also look at some recent patterns concerning urban specialization in China.

22.4.1.1.1 What big cities do and their skill composition

Figure 22.7 shows different manufacturing industries and their elasticities of local employment with respect to the metropolitan area population. The figure tells us two things. First, traditional industries producing standardized products such as wood products, furniture, and paper products have low elasticities, consistent with findings in the earlier work noted above. Higher technology industries such as the computer and

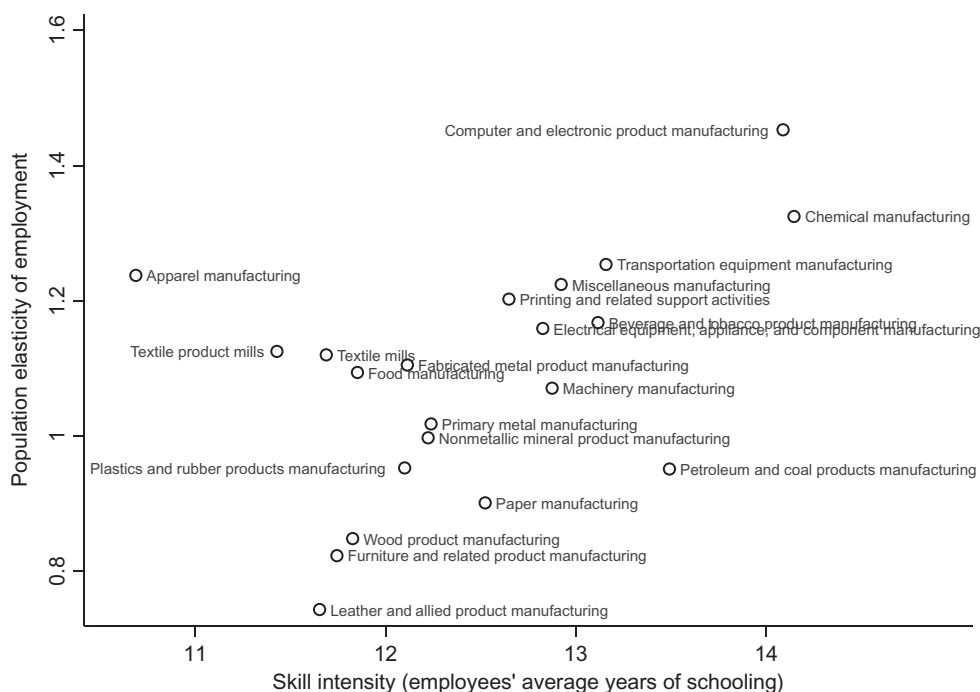


Figure 22.7 Relative industry composition position in the urban hierarchy and relationship to industry skill intensity. From Davis and Dingel (2013, Figure 14).

electronics industries have higher elasticities, consistent with the idea that high-tech production benefits from the diverse environments of large cities. Second, in general the skill intensity of industries is correlated with these elasticities, suggesting skill intensity also rises with city size (as [Davis and Dingel, 2013](#) show separately). Skill intensity is measured by the average number of years of schooling of people working in an industry nationally. The only strong outlier is apparel which is a low skill industry and which has a high elasticity. This may reflect the recent surge in the immigrant proportion of the workforce in the apparel industry, where migrants' first landing points are disproportionately cities such as New York and Los Angeles. [Figure 22.7](#) covers only manufacturing.

[Figure 22.8](#) looks at the role of big cities for the universe of workers, focusing on occupational mix. Traditional occupations such as farming, food preparation, and health care support have again lower city size elasticities and low education, while computing and mathematical, architectural, and engineering occupations have higher elasticities and high skill levels. Taken together, the findings show that today the United States has lower-skilled workers in standardized manufacturing and services in smaller cities, with higher-skilled workers in often more innovative and creative industries and occupations in larger cities.

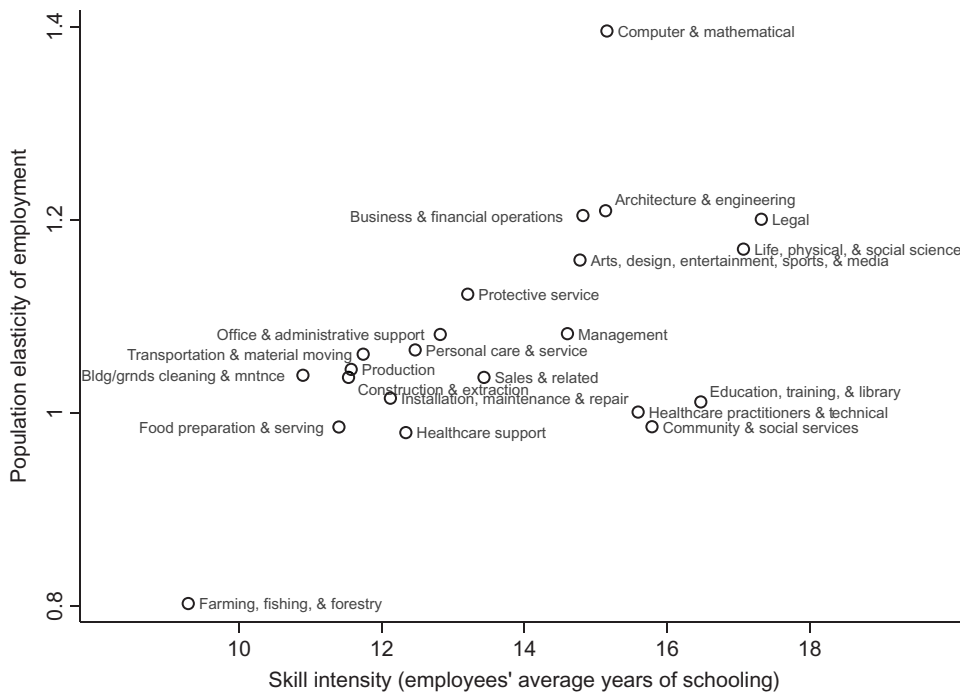


Figure 22.8 Relative industry occupational position in the urban hierarchy and relationship to industry skill intensity.

22.4.1.1.2 Specialization in the urban hierarchy

The next feature concerns the degree to which cities are specialized. For individual cities, the standard measure of the degree to which a city is specialized is the “Gini” specialization index taken from [Krugman \(1992\)](#):

$$K_j = \frac{1}{2} \sum_i^n |s_{ij} - s_i|,$$

where s_{ij} is subindustry i 's share of city j 's total industry employment, and s_i is industry i 's share of national total industry employment. The higher the index, the more specialized (less diverse) the locality is. The range is from 0, where the city's shares of different subindustries perfectly mimic the nation's share of different industries, to values approaching 1 for a city that is completely specialized in a minor product nationally. An alternative index takes the squares of the deviations, thus giving more weight to bigger deviations. However, that index is mechanically affected by the count of industries in the SIC, which changes over time (the index falls mechanically as the number of industries rises). The Krugman Gini coefficient is free of that mechanical effect.

The second and third columns in [Table 22.2](#) show the Gini coefficient for different size classes of cities for 1977 and 1997 based on County Business Patterns data. Two things are apparent for the United States. First, going down the urban hierarchy by city size class, we find specialization increases sharply. Second, the specialization index has declined for size categories over time, consistent with the manufacturing diffusion analysis in [Section 22.3](#).

The next columns in [Table 22.2](#) deal with a different aspect of activity composition across the urban hierarchy: how firms organize their internal functions by size class, leading to functional specialization by firms across the urban hierarchy. [Duranton and Puga \(2005\)](#) calculate the average number of executives and managers relative to production workers in manufacturing in all cities for 1950 and for 1990. Then they calculate that number within each size class and show the percent deviation of the size class in that year

Table 22.2 Specialization in manufacturing and function specialization across size classes of cities

Population (millions)	Sectoral specialization (Gini coefficient)		Functional specialization (management vs. production)	
	1977	1997	1950	1990
> 5	0.377	0.374	+10.2%	+39.0%
1.5–5	0.366	0.362	+0.3%	+25.7%
0.5–1.5	0.397	0.382	–10.9%	–2.1%
0.25–0.5	0.409	0.376	–9.2%	–14.2%
0.075–0.25	0.467	0.410	–2.1%	–20.7%
< 0.075	0.693	0.641	–4.0%	–49.5%

Source: [Duranton and Puga \(2005\)](#).

from the national average. Bigger cities have relatively more managers and executives in both years, but the degree of differentiation of managers and executive shares between small and large cities has increased enormously over time. Underlying this is a large increase over time in functional specialization by firms nationally (Kim, 1999), with production-oriented versus management-oriented activity increasingly in different locations.

Several articles explore functional specialization by firms across the urban hierarchy in recent years using micro data. Ono (2003) shows that in bigger cities, production plants found there rely more on the headquarters to buy business services for them. More generally, the headquarters are located in larger cities and enjoy a high degree of local scale externalities (Davis and Henderson, 2008). The headquarters outsource most services functions and are located in larger service-oriented cities (Aarland et al., 2007). Smaller cities house plants of firms in certain sectors of manufacturing and have relatively less business and financial services.

There are other dimensions to urban specialization and the hierarchy. A notable article by Fafchamps and Shilpi (2005) documents a pattern in specialization that may be typical in developing countries, using data from the Nepal Labour Force Survey. Note that countries at Nepal's stage of development have yet to develop a sophisticated manufacturing structure, let alone a corporate structure to produce. The data show how individuals allocate their hours to wage work, nonfarm self-employment, agriculture, construction, food processing, handicrafts, and other work. It also lists the main occupation of an individual for 56 International SIC occupation codes. Looking at patterns across 719 wards covering most of Nepal, the study authors have two key findings. The first concerns specialization in the allocation of time by individuals, which reveals a new result for the literature. Such specialization declines as people in a ward are less exposed to nearby urban populations, or live in less urbanized areas. An increase in the urban population nearer to a person induces more specialization in individual tasks—more Adam Smith specialization. The second finding concerns specialization at the ward level by the local population, where, as people in a ward are less exposed to nearby urban populations, ward specialization increases. This pattern suggests that wards nearer to cities can support a more diverse set of individual occupations while remote places are more specialized, paralleling at a different spatial scale what we saw in the second and third columns in Table 22.2. In contrast, Adam Smith specialization for individuals increases with greater exposure to urban markets.

22.4.1.2 Modeling the urban hierarchy

22.4.1.2.1 Initial attempts at a hierarchy: A benchmark model of city sizes and hierarchies The initial systems of cities literature (Henderson, 1974) envisioned an equilibrium system with differing types and corresponding sizes of cities, where each type was specialized in the production of one traded good. The idea was that if scale effects were ones of localization (internal to the own industry), specialization accentuated the efficiency gains from

agglomeration relative to the congestion costs of increasing city size. The model has workers with identical skills and capital. The equilibria are free-mobility ones, meaning that workers are free to move across cities and in equilibrium earn equal utility everywhere. Henderson considered two regimes. In the first regime, there are agents who operate in national land markets to set up cities, such as developers or city governments. City sizes under such a regime are determined by developers or local governments which set sizes to optimize local net rents or per worker utility. Equilibrium sizes represent a trade-off between the marginal benefits of enhanced scale effects in production versus the marginal costs of increased commuting or generalized congestion from increasing city size. In an urban hierarchy, bigger types of cities are specialized in industries with greater marginal benefits of enhancing scale. In the second regime, there are no organizing agents operating in national land markets, and cities form through what is now called self-organization, a term introduced by [Krugman \(1996\)](#). With self-organization workers move across cities to equalize private marginal products but individually cannot act to internalize externalities. As we will see, in this regime, there are a continuum of potential equilibria where cities are generally too large.

Initial extensions of the basic model focused on modeling the microfoundations of local scale economies, which in Henderson are presented as traditional black-box scale externalities. [Fujita and Ogawa \(1982\)](#) model information spillovers as decaying with distance, which provides an incentive for people to cluster in agglomerations. [Helsley and Strange \(1990\)](#) model how the efficiency of search and matching in labor markets improves in thicker markets. Building on earlier work developing monopolistic competition models in urban economics,² [Abdel-Rahman and Fujita \(1993\)](#) model diversity of local nontraded intermediates which increases with urban scale, and thus provides greater choice and efficiency for final good producers in the city. [Duranton and Puga \(2004\)](#) present a detailed analysis of how to add other, more behavioral elements when thinking about microfoundations of scale externalities.

Another set of initial extensions focused on two aspects of urban hierarchies. First, rather than cities being specialized in one export good, in reality most cities export more than one good. Second, as we have seen, diversity of traded good production (i.e., manufacturers) increases as we move up the urban hierarchy. An early attempt to add such elements was by [Abdel-Rahman and Fujita \(1993\)](#), who looked at a world with two cities and two final traded goods (which can be produced with diversified intermediate nontraded inputs). Each final good requires fixed inputs, where the requirement is lower for one good than for the other. However, if the two industries colocate, these fixed costs can be reduced for firms in each sector. In their equilibrium, there is a city specialized in the good with the lower fixed costs, and the other, potentially larger city is diversified. [Tabuchi and Thisse \(2011\)](#) present a similar model and outcome, but now the two goods

² See, for example, the special issue of *Regional Science and Urban Economics* in 1988 edited by Fujita and Rivera-Batiz.

differ in the unit costs of intercity trade. In this case, the specialized city is the one with the lower unit trade costs. While these models do give specialized versus diversified cities, the environment is not rich. The number of cities is exogenously set at two and the distinction between goods has limited intuition.

In the recent literature, much more sophisticated modeling of production and labor force structure across the urban hierarchy has emerged. We turn to these in the next section, but as a reference point, we review key aspects of the basic model reviewed in detail in both [Duranton and Puga \(2004\)](#) and [Abdel-Rahman and Anas \(2004\)](#).

A benchmark model. For our benchmark, we use as microfoundations for scale externalities the diversity of intermediate inputs framework. It is straightforward to reformulate the model to allow the scale benefits to be other microfoundations, such as greater scale promoting greater specialization of workers in their tasks ([Becker and Henderson, 2000](#); [Duranton and Puga, 2004](#)). A city has production functions for final and intermediate producers, respectively, of

$$Y = \left(\int_0^m x(h)^{1/(1+\varepsilon)} dh \right)^{1+\varepsilon}$$

and

$$X(h) = \beta l(h) - \alpha,$$

where $l(h)$ is labor input for firm h , and $x(h)$ and $X(h)$ are, respectively, inputs of type h for a final good firm and output of the intermediate good producer of type h . For other notation m is the endogenous number of intermediate good producers, L is the effective city labor force, Y is total final good output, the price of the final good is the numéraire, and the price of intermediate inputs is q . Using key results from standard cost minimization of final producers and from profit maximization and competition among intermediate producers,³ we find the reduced form expressions for final good output per worker in the city and wages, respectively, are

$$Y/L = CL^\varepsilon$$

and

$$w = (\beta/(1+\varepsilon))m^\varepsilon = (\varepsilon/\alpha)^\varepsilon (\beta/(1+\varepsilon))^{1+\varepsilon} L^\varepsilon.$$

Both output per worker and wages increase with city scale, as ultimately measured by total effective employment. Note the reduced form specification looks like black-box externalities.

³ For cost minimization, we have that the direct elasticity of derived demand is approximated by $-(1+\varepsilon)/\varepsilon$ and that final price $1 = (\int_0^m q(h)^{-1/\varepsilon} dh)^{-\varepsilon} = qm^{-\varepsilon}$, where the last term emerges in the symmetric equilibrium. Profit maximization conditions by intermediate producers and free entry allow us to solve for the wage level $w = \beta/(1+\varepsilon)q$, firm output $X = \alpha/\varepsilon$, and the number of such producers in the city $m = \beta\varepsilon/((1+\varepsilon)\alpha)L$.

Given these positive benefits of increasing scale, what economic forces serve to limit city sizes and serve as a counterbalance to scale benefits from agglomeration? To answer this requires the introduction of sources of urban diseconomies. Such diseconomies are typically modeled as coming from increases in urban commuting costs. The standard approach assumes a monocentric city with fixed lot sizes where all production occurs at a point in the city center. Following the specifics in [Duranton and Puga \(2004\)](#) for a linear city, each worker is endowed with one unit of time and the working time is $1 - 4\tau u$, where u is the distance from the city center and 4τ is the unit commuting costs. It is then easy to derive expressions for the effective labor force L , for total rents in the city, and for the net wage after rents and commuting costs, all as functions of the city population N .⁴ For use below we have

$$L = N(1 - \tau N); \text{ net wage} = w(1 - 2\tau N); \text{ total rents} = w\tau N^2.$$

The final step is to introduce the mechanism to determine city sizes. The standard one following the first regime in [Henderson \(1974\)](#) assumes the existence of “large agents” operating in national land markets who serve to coordinate agglomeration. These could be developers who own city land and set city sizes and any subsidies to workers or firms to maximize their profits, or alternatively (and equivalently) they could be city governments, who can tax away land rent income from landowners and set city sizes to maximize real income per worker. As an example, developers seek to maximize

$$\begin{aligned} \text{Profits} = \text{total rents} - \text{worker subsidies} &= w\tau N^2 - sN = \left(\frac{\varepsilon}{\alpha}\right)^\varepsilon \left(\frac{\beta}{1+\varepsilon}\right)^{1+\varepsilon} \tau N^{2+\varepsilon} (1-\tau N)^\varepsilon - sN \\ \text{subject to } \bar{y} &= \left(\frac{\varepsilon}{\alpha}\right)^\varepsilon \left(\frac{\beta}{1+\varepsilon}\right)^{1+\varepsilon} N^\varepsilon (1-\tau N)^\varepsilon (1-2\tau N) + s, \end{aligned}$$

where s is any subsidy developers pay workers to join their city and $\bar{y}$ is the going real income available for workers in national labor markets, as perceived by any city. These subsidies could also go to firms, but in this simple example this is irrelevant.⁵ Assuming that developers maximize profits with respect to s and N and that, with competition, cities earn zero profits, solving the problem gives the equilibrium (and efficient) city size⁶:

⁴ The population comes from integrating over the two halves of the city, each of length $N/2$. The rent gradient is derived by equating rent plus commuting costs for a person at u with that of a person at the city edge, where rents are 0. Total rents come from integrating over the rent gradient.

⁵ There is no misallocation of resources here, despite fixed costs of production and monopolistic competition, because diversified inputs are the only factor of production and enter symmetrically.

⁶ There is also the Henry George theorem where all rents in the city are paid out to workers in subsidies to cover the marginal externalities they generate (more workers bring more varieties and greater efficiency of final good producers). In particular, $dY/dN = (1 + \varepsilon)[(\varepsilon/\alpha)^\varepsilon (\beta/(1+\varepsilon))^{1+\varepsilon} N^\varepsilon (1-\tau N)^\varepsilon (1-2\tau N)]$, where the term in the square brackets is the private benefit of adding a worker (his/her net wage) and ε times the expression in square brackets is the externality, which also equals s in equilibrium.

$$N^* = \frac{\varepsilon}{\tau(1+2\varepsilon)}; \quad \partial N^*/\partial \tau < 0, \quad \partial N/\partial \varepsilon > 0.$$

As constructed, this is also the size that maximizes net income per worker, γ , including the subsidy set equal to average land rents.⁷ This implies that γ is an inverted-U-shaped function of N with equilibrium and optimum city size at this maximum. That equilibrium and optimal size coincide in this context depends on the use of subsidies to residents to effectively internalize scale externalities, as financed by land rents. If, for example, land rents go to absentee owners, as reviewed in [Abdel-Rahman and Anas \(2004\)](#), cities will be too small.

There are some loose ends before proceeding to recent developments. What happens under the self-organization regime? The requirement for a Nash equilibrium in worker location choices is that no worker wants to change cities in equilibrium. Given that income, γ , is an inverted-U-shaped function of city size, this has two implications. The first is that the equilibrium size is at the peak or to the right of the peak where $d\gamma/dN < 0$. That is, if a worker moves to another city (by increasing its size), he/she would earn less than what he/she earned in the city he/she left (where real income would rise as he/she left). Thus, it is also the case that cities to the left of the peak where $d\gamma/dN > 0$ cannot be Nash equilibria. The second implication is that all cities be of the same size so as to equalize real incomes. There is then a continuum of equilibria in city sizes between the peak and a size to the right of the peak, $N_{\max}$, where $\gamma(N_{\max}) = \gamma(N; N = 1)$. Beyond $N_{\max}$ workers would deviate to form a city of size 1, which would then induce migration flows and self-reorganization until there was a new equilibrium where all cities again had a common size between N^* and $N_{\max}$.

Thus, in general, city sizes under self-organization are oversized, potentially enormously so. However, there are models where under self-organization there are unique and more reasonable city size solutions. In the absence of optimizing city land developers, [Henderson and Venables \(2009\)](#) show that in a world with durable housing capital as a commitment device equilibrium city sizes are unique and that, while cities are oversized, they are only modestly so. [Behrens et al. \(2014\)](#) have another, reasonable self-organization equilibrium for the special case they focus on with a continuum of heterogeneous workers. However, most of the literature avoids the self-organization paradigm by assuming either that the number of cities is fixed so city formation is not an issue or that, with an endogenous number of cities, there are land developers who act as optimizing agents to achieve potentially efficient and unique outcomes.

Finally, as alluded to above, to get a hierarchy we would specify that there is a variety of final consumer products, or sectors, each produced with different degrees of scale economies (ε) in their use of local nontraded intermediate inputs. Having different ε is generally enough to guarantee specialization and a hierarchy, but that is fully assured

⁷ That is, N^* maximizes $[(\varepsilon/\alpha)^\varepsilon (\beta/(1+\varepsilon))^{1+\varepsilon} N^\varepsilon (1-\tau N)^\varepsilon (1-\tau N)]$ such that $d\gamma/dN = 0$ and $d^2\gamma/dN^2 < 0$.

if we also assume that the production of inputs is specific to each final good sector. As noted earlier, this assumption of “localization” economies means that there are no benefits to industries from colocating. With costs on the commuting consumption side (higher commuting distances and rents), specialized cities are more efficient than diversified ones, as they more fully exploit localization economies. With a fixed set of final goods, in a developer regime, we will have different types of cities, each specialized in one type of product as in [Henderson \(1974\)](#). The sizes of a city by type increase as ε increases across types. While here specialization involves final goods that are uniform in quality, in many recent applications, they could be diversified products within sectors (or types of cities) sold under monopolistic competition. We could also have each city specialized in one particular variety of a traded product Y under monopolistic competition with different values of ε , as in [Au and Henderson \(2006a\)](#), who estimate a simple structural model applied to China.

22.4.1.2.2 The second generation of hierarchy models

With this simple benchmark in mind, we now turn to the second generation of models developed in the early years of the twenty-first century by Duranton and Puga.

Nursery cities and the product cycle. The second generation of hierarchy models starts with [Duranton and Puga \(2001\)](#), who have an endogenous number of cities, introduce at least one type of diverse city, and develop models that relate to the larger economics literature. In their 2001 article, they build upon the product cycle hypothesis from international trade. That model seeks to explain why product innovations are carried out in major centers (in our case big cities), but, once standardized, production moves to lower-cost sites (in our case smaller cities). In [Duranton and Puga \(2001\)](#), there are m types of final goods, each produced by firms using varieties of type-specific diversified intermediate nontraded inputs. Diversified nontraded inputs of type j must be produced by workers with the same labeled aptitude, where there are thus m types of workers. Final good firms are subject to a probability δ of dying in a period, so there is firm turnover, with new firms appearing in each period. Most critically, each new firm draws an ideal type of intermediate input it must use, but it does not know what that type is. It experiments with different intermediate inputs of type j produced by workers with aptitude j , producing prototypes at a higher cost until it finds its ideal type. Once it chances upon its ideal type, its costs of production fall (thus signaling that the producer has found its ideal type).

How does this fit into urban structure? Using the developer framework for how cities are established, in equilibrium there are specialized cities, where for type j there are only workers and intermediate producers of type j in the city and all final good producers in the city have discovered their ideal type is j . For those specialized cities, scale benefits arise only from having more type j intermediate producers. Thus, as in the previous

subsection, specialization comes from maximizing scale benefits relative to commuting costs, by having only type j producers, given an absence of any cross-industry scale effects.

The second type of city is novel: a diversified nursery city. In such a city, all sectors are represented and there are equal numbers of each type of worker and of each type of intermediate good producer. Final good producers produce prototypes as they seek to learn their ideal technology. Why does this experimentation occur in diversified cities, rather than in specialized ones? [Duranton and Puga \(2001\)](#) assume that to move from city to city is costly; a final producer loses a period of production. Thus, to experiment by visiting different specialized cities can become quite costly, whereas to shift input types to experiment within the same city is costless.

Note two key aspects of the nursery city equilibrium just portrayed. The cost of moving across cities (loss of production for a period) must be sufficiently high relative to the scale economies from being in a specialized city, so new firms do not experiment in only specialized cities. But it cannot be so high that once firms know their ideal type they do not want to move to a specialized city (with its lower production costs) for the horizon of their life. Note that this tension also places limits on how relatively high the probability of dying may be. The triumph of the model is not just having a new type of diversified city, but in also formalizing an urban version of the product cycle model. Recently, empirical work and some theoretical work have focused more directly on the role of innovation in cities; this work is reviewed in the chapter by [Carlino and Kerr \(2015\)](#) in this handbook.

Functional specialization and diversity. [Duranton and Puga \(2005\)](#) explore a different type of hierarchy where rather than distinguishing only between product types, they also distinguish between functions. Production units of a firm use intermediate physical inputs and service inputs provided by their headquarters. The headquarters produces these services with intermediate service inputs and labor. Both services and physical intermediate inputs are produced with labor and are not tradable across cities. As in the nursery city model, there are workers belonging to specific occupations (aptitudes) and thus firms in different sectors. Production units use sector-specific intermediate inputs. In contrast, the headquarters of different firms in any city use a common set of business service inputs. So all types of headquarters use lawyers and accountants, but only apparel firms use textile inputs. Firms may spatially integrate so the headquarters and production are located in the same city or they may be multilocation firms, with their headquarters and production units in different cities. Most critically to get their results, multilocation production raises the cost of a production unit to acquire its headquarter services by a factor $\rho > 1$, relative to it being in the same location. However, having the headquarters in separate specialized business service cities allows a greater diversity of intermediate business services of benefit to all types of firms and their headquarters.

Given these implicit trade-offs, the equilibrium has a multilocation pattern for firms, and there are two sets of cities. One set comprises cities specialized in headquarter and business service production. The other set comprises cities specialized in the production

of one type of final good and their corresponding intermediate inputs. [Duranton and Puga \(2005\)](#) call this functional specialization by cities, where now the diversified city is one where the headquarters of different production sectors enjoy a diversity of common business service inputs that are not traded across cities. We note this functional specialization equilibrium will not exist if the cost of having production units acquire headquarter services from other cities is high enough.⁸

22.4.1.2.3 The third-generation models

In the last few years, several articles have introduced more sophisticated considerations into modeling urban hierarchies. Prior work, even in the second-generation models, took a simple approach to looking at urban specialization and diversity. Very recent work has introduced several innovations. First and foremost is allowing for labor heterogeneity, not just different labor types (horizontal differentiation) but also different labor talents or skills (vertical differentiation). This introduces the possibility of labor sorting by talent across the urban hierarchy. Second, in bigger cities, competition among firms may be “tougher” and different qualities of firms may survive. Third, there may exist more complex sorting by industries across the urban hierarchy, based on more complex specifications of interindustry interactions and scale externalities.

Such sorting is critical to the evaluation of urban productivity. In developing and even developed countries, some policy makers evaluate that bigger cities are inherently more productive. That has in certain instances become a basis for advocating that these cities should be effectively subsidized at the expense of smaller cities, an issue we will return to in [Section 22.4.3.2](#). However, small cities persist in developed market economies, suggesting that they are competitive and thus productive. The issue is that in the data we typically observe higher measured output per worker in bigger cities, which could be a basis for the evaluation of policy makers. But this does not mean that bigger cities are more productive. The puzzle can be explained by the types of sorting just noted.

First, we know from [Figures 22.7 and 22.8](#) that more educated and higher-skilled workers sort into bigger cities. So if we observe higher output per worker in a bigger city, the question is to what extent is that because of pure productivity effects versus because of higher quality labor. Models that tackle sorting across cities help us to understand that issue better. Second, if competition in bigger cities is tougher so that only higher productivity firms survive there, that also lowers the component of higher output per worker in bigger cities owing to pure productivity effects. Third, there is industry sorting across cities, where only certain types of industries are found in bigger cities. In the early part of this section, we discussed the idea that industries with greater

⁸ If ρ exceeds a critical value, then the equilibrium has only integrated production. Then each city type specializes in production of one type of final output and hosts just the headquarters of the firms in that city and their corresponding intermediate physical and business service suppliers.

localization economies of scale should be found in larger types of cities, with also higher costs of living, while those with lower localization economies may be found in smaller types of cities. Equilibrium in national labor markets with equalized real wages will also require higher output per worker and wages in bigger cities to offset higher costs of living in those cities. The key is that different sizes of cities house different industries; or smaller cities are competitive in what they produce. However, recent work suggests that the issue is more complicated when there are cross-industry externalities. Maybe an industry with high localization economies in a bigger city would really benefit from having an industry with lower localization economies colocate there, but that may not be realized in a market equilibrium and makes local policy enactment and evaluation of productivity more complicated.

There are several articles that tackle theoretical models of sorting of workers across cities, apart from empirical modeling, which we do not cover here (e.g., [Baum-Snow and Pavin, 2012](#)). The first article we look at examines sorting across cities, with the distinction that the model links such sorting across cities to residential sorting within cities ([Davis and Dingel, 2013](#)). To achieve this neat link, sorting in that article always goes in the direction of having more skilled workers sort into bigger cities. Another article, which focuses only on sorting across cities, questions the presumption that there is monotonic sorting ([Eeckhout et al., 2014](#)). The third article we discuss combines sorting of workers across cities with the idea that competition may be tougher in bigger cities ([Behrens et al., 2014](#)). This article has a number of nice innovations, one being the endogenous formation of firms within cities. Finally, we analyze the article by [Helsley and Strange \(2014\)](#) on sorting of industries across cities in the face of cross-industry scale externalities.

Sorting within and across cities. [Davis and Dingel \(2013\)](#) develop a model of sorting across and within cities, albeit in a context where the number of cities is set exogenously. Cities have internal space, which is required if workers are going to sort with regard to where to live within the city. Similar to the benchmark model, final output is produced just with intermediate inputs, but now from a fixed set of intermediate input sectors. In [Davis and Dingel \(2013\)](#), intermediate inputs are sold competitively, traded costlessly within and across cities, and produced by workers with different skills, where there is perfect substitutability among skills in production in any sector σ . The higher σ , the more “advanced” the sector is, as defined below. A worker living in city c at location δ in the city with skill ω chooses which sector σ to work in so as to maximize wages net of rent, or

$$\max_{\sigma} p(\sigma)A(c)D(\delta)H(\omega, \sigma) - r(\delta, c).$$

The worker takes the price $p(\sigma)$ of output in the sector as given. Locations in a city are ordered by values of δ , with the most desirable at $\delta = 0$, and $D' < 0$. While the interpretation can be quite general, to fix ideas and to meet a regularity condition that better locations be “relatively scarcer” in a smaller city, we adopt the Davis and Dingel example

where all cities are circular with fixed lot sizes, δ is the distance from the city center, and $D(\delta)$ is linear. Note furthermore that $r(\delta, c)$ is the rent at location δ in city c , and $A(c)$ is an urbanization productivity level in the city where, for $Lf(\omega, c)$ being the quantity of ω skilled people in city c , $A(c) = J\left(L \int_{\omega \in \Omega} j(\omega) f(\omega, c) d\omega\right)$, $J', j' > 0$. Heterogeneous individuals have density function $f(\omega)$ on support $\Omega \in [\underline{\omega}, \bar{\omega}]$. An equilibrium will have relatively more high-skilled people in bigger cities and thus A is higher in bigger cities because of both scale and skill composition. Finally, worker technology, $H(\omega, \sigma)$, increases with ω and is supermodular (Costinot, 1999), so H has a larger value for the same skill ω in a more advanced sector.

To solve for the within-city and across-city sorting, Davis and Dingel (2013) utilize the perfect substitutability of skill in production of intermediate inputs. Then, in equilibrium the marginal returns to ω in sector σ are independent of the assignment of ω 's to the sector. With perfect substitutability, the worker's choice of σ simplifies to $M(\omega) = \max_{\sigma} p(\sigma) H(\omega, \sigma)$ and defines $G(\omega) \equiv H(\omega, M(\omega)) p(M(\omega))$, $G' > 0$, where then a worker's choice of the sector in a city is independent of his/her location choice δ . This in turn yields a simplified location problem within the city of $\max_{\delta} A(c) D(\delta) G(\omega) - r(\delta, c)$. Within a city, higher-skilled people outbid lower-skilled people for better locations, because they have a higher willingness to pay for better locations, or $\frac{\partial^2}{\partial \delta \partial \omega} A(c) D(\delta) G(\omega) < 0$.

We can now turn to some properties of an illustrative equilibrium with two cities, where city c will be larger than city c' in equilibrium. If we think of $A(c) D(\delta)$ as measuring the attractiveness of a location in city c , then $A(c) > A(c')$ and $L(c) > L(c')$. Why? With rents standardized to zero at each city edge, those least desirable locations in each city in equilibrium must have the same general attractiveness as they will house the same type of worker, the lowest-skilled ($\underline{\omega}$) people. Across cities, in the larger city, the highest-skilled people will live nearest to the city center in locations that are more desirable than any in the smaller city. Only the very highest skilled people in $(\omega, \bar{\omega}]$ are found in the larger city living between $\delta(c) = 0$ and $\delta(c) = \bar{\delta}$. At $\delta(c) = \bar{\delta}$ and $\delta(c') = 0$ across the respective cities, workers have the same skills $\tilde{\omega}$ at those respective locations and pay the same rents. After that there are people of all lower skills in both cities. For a person of skill $\omega < \tilde{\omega}$ found in both cities, they must be equally well off, pay the same rents, and face the same $A(c) D(\delta)$. Thus, an equally skilled person in the bigger city will have a higher δ and a lower $D(\delta)$.

Given the assumption noted earlier of "relative scarcity" of better locations in smaller cities, Davis and Dingel (2013) are able to show that the bigger city not only houses an entire segment of the highest-skilled people but for skill groups in common it has relatively more higher-skilled people than lower-skilled people. Under that condition, employing results in Costinot (2009), they show that higher-skilled people are more likely to work in higher-quality sectors and that such sectors are disproportionately found

in the bigger city. These results become the basis for the empirical work reviewed earlier in this chapter on the greater intensity of higher-skilled types in larger cities. In summary, the model neatly shows the intertwining of skill and scale externalities in defining productivity differences between bigger and smaller cities, even allowing for skill differentials to contribute to scale externalities. And interestingly, it links sorting across cities to within-city sorting.

Eeckhout et al. (2014) present a related model which focuses just on sorting across cities. However, in their work, sorting need not be monotonic with ever-increasing skills in larger and larger cities. In their model, there are two cities with exogenous city-specific productivity differences and three skill types. Their innovation is to explore the effect of differing degrees of complementarity between skill types. They suggest that there is extreme skill complementarity, where the larger (higher inherent productivity) city has a greater representation of both the highest-skilled and lowest-skilled types. Intuitively, high-skilled workers consuming low-skill services enjoy benefits from mutual location; or office workers in Manhattan enjoy food carts.

Sorting and selection. In another ambitious article, Behrens et al. (2014) consider both sorting of skill types across cities and selection within cities as to which people become entrepreneurs versus stay as workers. Selection raises the issue of tougher competition in bigger cities and whether firms in bigger cities are inherently more productive. The output setup is as in the benchmark model with local intermediate inputs going into production of one final numéraire good sold competitively. Intermediate inputs are produced by an entrepreneur using labor, l , with output increasing with the productivity of the entrepreneur. Individuals are given a draw of talent, t , which they know upon the draw. Then when they choose (irreversibly) the city they will live in, they get a draw of luck, s . The distribution they draw s from is the same for each city. A person's potential productivity as an entrepreneur is $\psi = st$, with firm output being ψl . Intermediate goods are sold under monopolistic competition to final producers, so intermediate entrepreneurs earn profits, which is their compensation and which increase with their ψ .

A person decides whether to be a worker or an entrepreneur on the basis of ψ , where within a city there is a critical value of ψ . People with values below the critical value choose to be workers (with the same wage and productivity—workers get no benefits from higher talent or luck), and those with a value above the critical value are entrepreneurs. The selection issue is whether bigger cities then typically have better firms. What about sorting across cities? It turns out that the sorting and selection issues are interlinked. Although the details in the article are involved, the basic idea is that there are the scale effects as in the benchmark and the sorting benefits as in Davis and Dingel (2013). Bigger cities are more productive because they can house more varieties of intermediate inputs and these scale effects interact with talent. In a version of log-supermodularity, the

expected return to talent rises with city scale and cities with more talented people have higher marginal benefits of agglomeration for a given size to be traded off against urban diseconomies. While there is sorting, in equilibrium selection is unimportant: each city has the same proportion of workers to entrepreneurs. So although bigger cities have more talented firms, the relative cut is the same in each city regardless of the size.

In Behrens et al., free migration can lead to two types of equilibrium. There can be a symmetric equilibrium where all cities are identical if the variation in talent across people is limited. But the interesting equilibria are asymmetric ones. Behrens et al. do not have city developers, but rely on properties of their special case for an equilibrium to define city sizes that emerge through self-organization. In their special case, cities are talent homogeneous, rather than the usual case of each city having a segment of the talent distribution. Each city has only one talent t and city sizes increase as the talent level rises. Cities under self-organization are too large but not catastrophically so. Intuitively, if cities get too big, people with the intended talent will want to move to less talented cities. The uniqueness and existence of such equilibria are not guaranteed.

In summary, Behrens et al. (2014), similar to Davis and Dingel (2013), have talent or skill sorting across cities, with bigger cities being in part more productive simply because they have more talented workers. But they also deal with selection and the choice of people to become entrepreneurs, where the relative proportion of workers to firms is the same across cities of different sizes. In evaluating the sources of higher productivity in bigger cities, the modeling suggests that sorting by talent is a key source of higher output per worker, while selection is not. This finding parallels the empirical results in Baum-Snow and Pavin (2012).

Cross-fertilization in externalities and sorting of industries. The benchmark model assumes that industries are characterized by simple specifications of externalities, which potentially lead to urban specialization. In an innovative article, Helsley and Strange (2014) explore cross-fertilization among industries in a generalized fashion. They explore equilibria and optima where cities may be of mixed employment or nonspecialized, and there can be a hierarchy of the nonspecialized. The article focuses on how under self-organization specialized cities are not just oversized but may also have poor compositions of output. The article is summarized in the chapter by Behrens and Robert-Nicoud (2015) in this handbook. Some key results are as follows. Pareto efficient size cities where different city types have a plurality of own type population but an optimal mix of other worker types can never be stable equilibrium city compositions and sizes. In an equilibrium, own type workers will want to agglomerate further with own type workers in some types of cities, compared with the optimum. If it is efficient for all cities to specialize when complementarities are weak, we have the usual problem of equilibrium oversized specialized cities under self-organization. However, we can also have an equilibrium with specialized cities when the optimum would dictate mixing. And when we do have

mixing, compositions and sizes are inefficient. For policy makers, the issue of how to enact optimal or second-best policies and institutions is not covered in the article, but the challenge is there.⁹

22.4.2 Dynamics in the urban hierarchy

So far the models we have presented are basically static, or, at best, stationary-state models. To examine the rapid changes in urban systems which occur with economic development, we need models of the dynamics of the urban hierarchy. Such models are limited in number, and many were reviewed in earlier handbook chapters (Gabaix and Ioannides, 2004; Henderson, 2005). Given that, we take a different expositional approach in the first part of this subsection. We start by interweaving a discussion of empirical relationships about the dynamics of the urban hierarchy with a brief review of some key models which sought to explain certain patterns in the data. These have to do with growth of city sizes and numbers over time and the stability of the size distribution of cities over time. We then turn to a more recent focus on why production patterns of individual cities and cities across the urban hierarchy change with time. There we return to the format of presenting empirical relationships and then models that relate to them.

22.4.2.1 Facts and concepts concerning the size distribution of cities

22.4.2.1.1 Growth in city sizes

In general, over recent centuries cities have been growing in population size. To see this for a more recent time period for which there are good data, Henderson and Wang (2007) take all metropolitan areas in the world from 1960 to 2000, divide them by mean size in any time period to get a relative size distribution, and define a relative cutoff point so that the minimum size to mean size is stable over time. They call this the relative size distribution. Within this distribution, the absolute mean (and median) size doubled from 1960 to 2000 worldwide. There are two explanations for why city sizes have increased with time, given size is determined by a trade-off between scale economies and diseconomies.

The first is that scale economies have been increasing relative to diseconomies. Black and Henderson (1999a) present an endogenous growth model where economic growth is

⁹ We note two comments on the model. The first is based on the specification of cross-industry externalities. There are I types of workers where any type is employed in the corresponding industry of that type. Output per worker in sector i in city j is a function of the vector of the number of workers in different sectors in the city, $\mathbf{n}_j$, such that output per worker in sector i in city j is $g_i(\mathbf{n}_j)$. The comment is that the study authors assume not just that there are localization and urbanization economies where $\partial g_i / \partial n_{ij} > 0$ and $\partial g_i / \partial n_{kj} > 0$, but there are complementarities where $\partial^2 g_i / \partial n_{ij} \partial n_{kj} > 0$, which is a special assumption without empirical validation. Second, they only fully solve for self-organization equilibria. How equilibria would look with optimizing developers who can cross-subsidize worker types within different types of cities is less clear, as is what institutions or policies would be required to achieve optima.

fuelled by human capital (knowledge) accumulation. In that model, human capital externalities interact with scale economies at the local level to enhance overall economies of agglomeration. They correlate differential growth rates of US cities with differential growth rates of local human capital. [Rossi-Hansberg and Wright \(2007\)](#) performed similar modeling on growth of city sizes. Recent work by [Desmet and Rossi-Hansberg \(2009\)](#) has a more nuanced approach, involving the endogenous evolution of scale externalities as part of their work focusing on transition dynamics.

An alternative to increasing scale externalities as the explanation for increasing city sizes is that diseconomies have dissipated with technological progress. The Alonso–Muth model emphasizes the decline in commuting costs as a driver of city spread. But it could also be the basis of increasing city sizes, with an eye to the technological revolutions of the last 120 years—the development of transit systems, the invention of the automobile, and the construction of multilane high-speed highway systems with rays and rings for cities. Empirical work by [Duranton and Turner \(2012\)](#) supports this idea. Finally, if we think outside the traditional models, growth in human capital per person may be associated with better technologies and management techniques in planning of cities and in managing urban diseconomies.

Although we have two reasons for city sizes to be increasing, that does not mean there will be necessarily a shrinking number of cities. In [Black and Henderson \(1999a\)](#) and [Rossi-Hansberg and Wright \(2007\)](#), city numbers may also increase with national population growth, as long as the rate of national population growth exceeds the growth rate of individual city sizes. These articles assume a fully urbanized world. Growth in city numbers in developing countries is also driven by urbanization, or the move out of agriculture as discussed in [Section 22.2.1](#).

22.4.2.1.2 Stability of the relative city size distribution and size ranking of larger cities

City size distributions for countries are remarkably stable over time, and some argue that they are either globally ([Gabaix, 1999](#)) or locally ([Eeckhout, 2004](#); [Duranton, 2007](#)) approximated by a Pareto distribution and thus obey Zipf's law. [Henderson and Wang \(2007\)](#) illustrate this stability for the world size distribution from 1960 to 2000. [Black and Henderson \(2003\)](#) and [Harris-Dobkins and Ioannides \(2001\)](#) show this similarly for the United States over many decades. To be clear, these exercises look at just cities, not at the spatial transformation of the universe of space as described in [Section 22.3](#). Theoretical modeling pioneered by [Gabaix \(1999\)](#) and developed more fully by [Rossi-Hansberg and Wright \(2007\)](#) and [Duranton \(2007\)](#) argues that stochastic processes in particular contexts such as ones that obey Gibrat's law generate a stable size distribution of cities over time approximated by Zipf's law. A potential problem is that these models also have all cities transiting continuously through the size distribution of cities, in partial contrast to the next fact.

Evidence suggests that the biggest cities historically tend to remain the relatively biggest cities in a country over long periods of time. There is little move downward from the

top rung of cities in a country (Eaton and Eckstein, 1997; Black and Henderson, 1999b, 2003). Eaton and Eckstein (1997) show that the ranking of cities by size has been remarkably stable in France and Japan over the prior 100 years or more. In a Markov process based on 10 decades of data, Black and Henderson (1999b, 2003) show that mean first passage times for a US city in the top 5% of population size to transit to the bottom 35% is many centuries (which is a time horizon way out of sample). The question is why are big cities so slow to move down the size ranking? Glaeser and Gyourko (2005) and Henderson and Venables (2009) claim that city durable capital is an explanation for why big cities retain populations in the face of bad shocks and competition. Arthur (1990) and Rauch (1993) stress information externalities are embedded in place, where bigger cities have a large accumulated stock of knowledge that is not readily transferable.

22.4.2.2 Churning and movement of industries across the urban hierarchy

In this part, we start with some facts about the movement of industries across cities, which is also related to the movement within cities (from the core to the periphery). We then turn to a discussion of two recent relevant models.

22.4.2.2.1 Facts about industry movement

Churning is the process whereby cities over time lose their existing export industry or industries, to be replaced by different export industries. So an automobile city of today may become an electronics city in the next decade. Churning can be defined directly with a churning index based on the work of Davis and Haltiwanger (1998) and used in Duranton (2007), or can be based on mean first passage times in a Markov process (Black and Henderson, 1999b, 2003; Duranton, 2007). The mean first passage time for a top city industry to transit from the top 5% to the bottom cell of five cells is a small fraction of the mean first passage time for a city in the top 5% of the population to transit to the bottom cell.

While Duranton observes generalized churning in US and French data, there are other sets of empirical findings, more specific to the economic development process. A first set of findings concerns the degree of specialization of cities in the urban hierarchy. As noted in Table 22.2, the degree of specialization of cities in the United States has declined over the last 30 years. In contrast, South Korea, at a different stage of development, showed increasing specialization of cities from 1983 to 1993 for most industries, while diversity increased at the more aggregate regional level. So regional economies diversified but industry concentration at the city level increased (Henderson et al., 2001). Another example is China, where average specialization at all spatial scales increased from 1995 to 2008. Table 22.3 shows that specialization for both urban counties and rural counties increased from 1995 to 2008, also at the larger spatial scales of the prefecture and metropolitan area (city proper). Note that Chinese cities in general as a group are less specialized than the individual urban districts making up those cities, consistent with an idea that there is neighborhood clustering of like activities within cities.

The second set of empirical findings concerns the phenomenon in developing countries of industrial decentralization from the cores of the largest metropolitan areas. At the early stages of national economic development, modern manufacturing in a country may be largely confined to the core city of the largest metropolitan area(s) for reasons discussed below. This concentration is followed by two stages of decentralization: first out of the core to peripheries of metropolitan areas and then from metropolitan areas to hinterlands.

The idea is illustrated using data for Korea and China. For Korea, [Table 22.4](#) looks at the evolution of manufacturing shares within the national capital region of Seoul, Kyonggi province. While Seoul metropolitan area has retained a fairly constant share of the population in its local region, its share of manufacturing employment declined dramatically during the 1970s and the 1980s, starting at 76% in 1970 and declining to 30% by 1993. This is movement of industry out of Seoul to nearby satellite cities and ex-urban areas. [Table 22.4](#) also compares the evolution in just the 10 years from 1983 to 1993 of shares of national manufacturing employment held by the three main metropolitan areas in Korea, their satellite cities, and then the rest of the country. This is a second stage of decentralization where the three core metropolitan areas continue to lose share. The losses are no longer to the satellite cities, but are beyond, to the hinterlands. The

Table 22.3 Changing specialization in China (three-digit industry breakdown)

	1995 Krugman Gini coefficient (manufacturing)		2008 Krugman Gini coefficient (manufacturing)	
	Mean	Median	Mean	Median
Prefecture	0.4033	0.3978	0.4694	0.4741
City proper (urban districts, 2010)	0.3059	0.2863	0.3525	0.3460
County (rural units, 2010)	0.4218	0.4185	0.4612	0.4574
County (urban units, 2010)	0.4359	0.4294	0.4825	0.4749

Source: Authors' own calculations, based on about 150 three-digit industries in each year which show positive employment.

Table 22.4 Stages of decentralization in Korea

Share of Seoul in Kyonggi province (National Capital Region)

	1970	1980	1983	1993
Population	62%	63%	67%	61%
Manufacturing Employment	76%	61%	45%	30%

Share of national manufacturing employment

	1983	1993
Seoul, Pusan, and Taegu metro areas	44%	28%
Satellite cities of Seoul, Pusan, and Taegu metro areas	30%	30%
Other cities, rural areas	26%	42%

Source: [Henderson et al. \(2001\)](#) and related calculations.

hinterlands' share rose from 26% to 42% in 10 years, at a time when their population share declined modestly. This shift to the hinterlands is correlated with the extensive investment in highways and telecommunications Korea undertook in the early 1980s to service hinterland areas. The overall dispersion of manufacturing is also consistent with manufacturing becoming a mature industry, as discussed in [Section 22.3](#).

For China, [Table 22.5](#) shows the decline in the shares of areas defined as core urban counties of metropolitan areas in 1990 in national manufacturing employment from 1995 to 2008. New urban counties are on the periphery of these 1990 urban cores, or are the new suburbs. Their employment shares more than double. But hinterland towns labeled as county towns also see a modest rise in their shares. Note the high concentration of services in the original urban cores in 2008 (we do not know the 1995 numbers for services), far in excess of their shares of either population or manufacturing employment.

[Desmet et al. \(2015\)](#) show a corresponding trend in India, looking at the growth of manufacturing versus service employment in districts where they are initially concentrated versus in districts where they are not. The time period is short, 2000–2005, but still the patterns are striking. As illustrated in [Figure 22.9](#), they fit locally a trend with error bands, and, as the trend moves to higher-density districts with fewer observations, the error bands widen. For manufacturing, there is strong mean reversion whereby districts with high densities in 2000 grow much more slowly than districts with low densities in 2000. The pattern for services is quite different. High-density districts on average have higher growth rates than at least the middle-density districts. At the upper end, growth rises with density. Overall, this suggests decentralization of manufacturing as in the Korean and Chinese cases, while services are concentrating even more in the high-concentration districts found in the biggest cities.

22.4.2.2.2 Modeling industry movement across cities

The facts presented in the previous section concern churning and the general turnover of industries in cities, changes in the degree of specialization in the urban hierarchy, and

Table 22.5 Stages of decentralization in China

Share of "nation"	Industry		Services	Population
	Share of national industry employment		Share of national services employment	Share of national population
	1995	2008	2008	2000
Urban counties in both 1990 and 2000	52%	41%	58%	28%
New urban counties	6.2%	13%	6.3%	5.5%
County towns	18%	22%	12%	18%
Other rural counties in 2010	24%	24%	24%	48%

Source: Authors' own calculations.

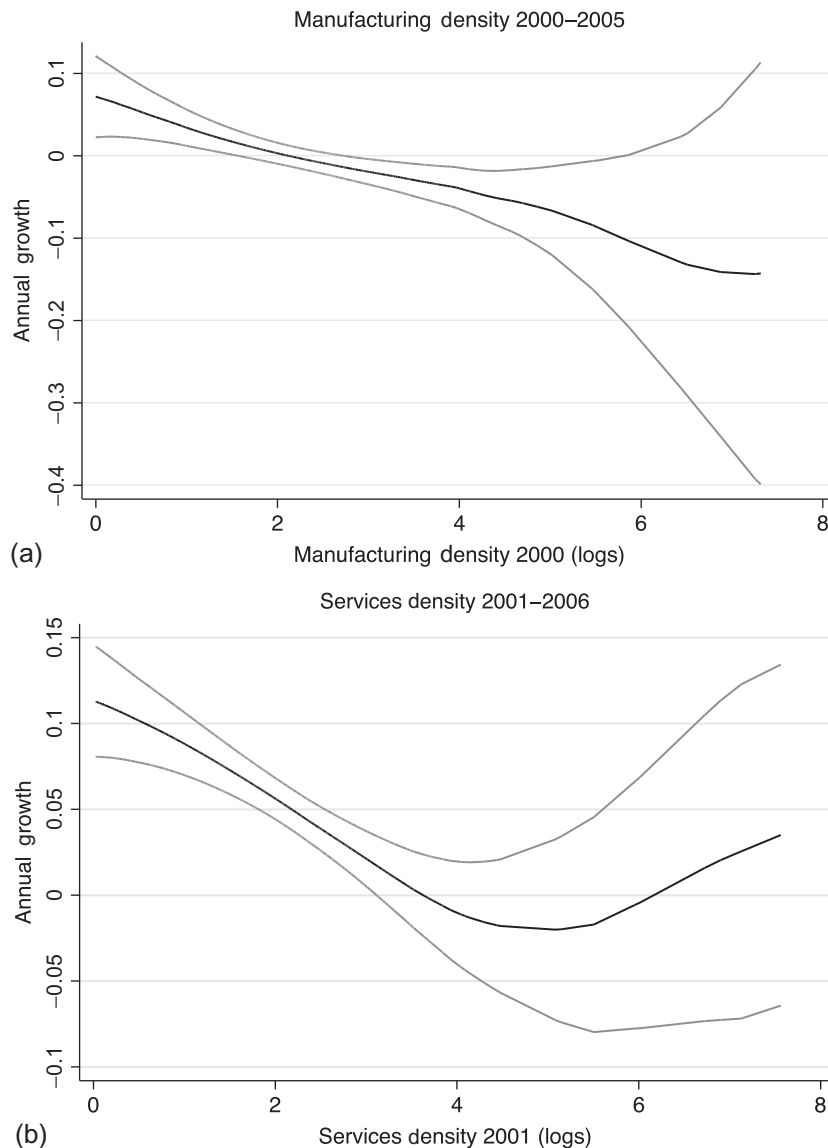


Figure 22.9 Decentralization of manufacturing and centralization of services in India: (a) manufacturing and (b) services. Source: (a, b) [Desmet et al. \(2015\)](#).

patterns of industry movements across the urban hierarchy. We outline a model that deals with churning *per se* and then review a variety of relevant models that deal with industry movements across the hierarchy.

Churning. [Duranton \(2007\)](#) adapts the Grossman–Helpman quality ladder model to an urban setting, with the aim of presenting a model of the following facts. Cities are slow to

change their relative sizes; the overall size distribution of cities is remarkably stable; and industries move quickly across cities, with cities changing production patterns. In the Grossman–Helpman quality ladder model, there are a fixed set of consumer goods, but they can be produced with ever-increasing quality. Quality is a ladder process where there is one current best quality $\bar{j}(z, t)$ for sector z at time t . Research by research firms is ongoing to improve that quality. Eventually that research leads in sector z to an advance discovered by one research firm. That firm then gets the (nontransferable) patent to produce that product and focuses on production activities, enjoying monopoly rents in production in industry z until there is the next move up the ladder. Only the research firm that discovers the latest quality level produces the product, pricing so as to exclude potential lower-quality producers.

Duranton adds an urban component. He assumes a fixed given number of cities, each specialized in the production of a different first-nature good, assumptions which anchor cities so that none can disappear, or become unpopulated. The action lies in second-nature goods which involve innovation and are completely footloose. Duranton makes two key assumptions. Production must occur in the place where a winning research firm makes a discovery. Production requires information from the research firm which can be transmitted only locally, such as through hiring the former research workers of the winning firm to be involved in production. Second, in order to be productive, all research firms focused on innovating in z must be located in the current city where z is produced. If all innovations, as in Grossman–Helpman quality ladder model, are within the own industry, then production would never move. Duranton introduces cross-industry innovation. The cumulated expenditure $\lambda^k(z)$ by research firm k focused on innovation z has a probability $\beta\lambda^k(z)$ of inducing a winning move up the quality ladder in industry z , but also a probability $\gamma\lambda^k(z)$, $\gamma < \beta$, of inducing a winning innovation in industry z' . The probability of an innovation in industry z is $\beta\lambda(z) + \gamma\sum_{z \neq z'} \lambda(z')$, where $\lambda(z)$ is the cumulated expenditures of all research firms focused on innovation in z . If a research firm working on z happens to make a winning innovation in z' , the production of z' moves to the city where this firm is located, generating churning.

In the steady state, there are several key results. First, there is industry churning: the location of production for second-nature products with footloose production will change over time driven by cross-industry innovation.¹⁰ Second, the innovation process leads to a stable size distribution of cities that locally approximates Zipf's law, so the size distribution of cities remains time invariant. Third, however, there is motion for individual cities. Bigger cities which (by accident) have accumulated innovations and production will lose and gain sectors over time, but the net result will be mean reversion, with

¹⁰ An older heuristic version of this was that traditional producers and their nearby research firms become “complacent” and the innovation occurs in new locations.

the biggest cities growing more slowly (or losing employment) relative to smaller cities. So there is a transition process where small cities move up and bigger cities move down (slowly).

Explaining patterns of movement across the urban hierarchy as economic development proceeds. We have two sets of shifts across the urban hierarchy. The first is the changing degree of specialization of cities and regions. The second is the movement of industry out of the core of the largest metropolitan area(s) to peripheries and then to hinterlands. How might we explain these shifts? In [Section 22.4.3](#), we will argue that public policy, transport investments, and innovations may play a role in explaining both of them. Here we focus on modeling that involves changes in production technology. In [Desmet and Rossi-Hansberg \(2009\)](#), as reviewed in [Section 22.3](#), two waves of GPT—electricity and IT—induced first the concentration of initially more dispersed manufacturing into high-density locations in the United States, and then several decades later the deconcentration of manufacturing from the most densely populated areas, to be replaced by services. Both also involve changing specialization at a more aggregate level. Correspondingly, we might think of developing countries experiencing technology transfers and adaptation. Learning with adaptation of foreign technologies is initially efficiently concentrated in the densest locations; but, as manufacturing technologies are adapted and standardized, scale externalities may diminish and disperse. Manufacturing moves out of the most densely populated locations, to be replaced by services.

The ideas in [Desmet and Rossi-Hansberg \(2009\)](#) also relate to within-metropolitan-area deconcentration historically in the United States and in developing countries today. In the United States, there was the shift of manufacturing with electrification to continuous-process production in the early twentieth century, where continuous-process production requires single-story buildings and hence a lot of land. Land being much cheaper at the city fringe than in the center provides an incentive for manufacturing to relocate out of core cities as it did in the early and middle twentieth century in the United States. Related to this, if the service sector within a city starts to enjoy greater marginal local agglomeration benefits than manufacturing, that makes the service sector better able to outcompete manufacturing for high-priced land in city centers, a point further developed in [Desmet and Rossi-Hansberg \(2014a\)](#).¹¹

22.4.3 Policies affecting the spatial allocation of resources

Government policies and institutions strongly influence the structure of the urban hierarchy. There are a whole range of policies, such as those governing trade, minimum

¹¹ Related to this, in the [Fujita and Ogawa \(1982\)](#) model, a decline in manufacturing externalities (the value of information spillovers within the city as technology standardizes) leads to the formation of more urban centers away from the core, fostering the development of subcenters to which workers can commute more cheaply.

wages, capital markets, and fiscal decentralization, which in older work (Renaud, 1981; Henderson, 1988) as well as more recent work are recognized as affecting the allocation of resources across the urban hierarchy. For example, policies which affect the national composition of products then affect the sizes and numbers of cities producing products favored by trade policies. As such, these policies will differentially affect cities through the urban hierarchy. So if trade policies favor steel at the expense of textiles, the national composition of cities will change so that the relative number of cities engaged in steel production or inter-related products will increase. These may be bigger types of cities than those engaged in textile and related production such as apparel production. Minimum-wage policies which fix nominal wages may bite only in big cities with higher nominal wages but not higher real wages.

It is beyond the scope of this chapter to review all these policies. While many have been covered in older research, on some there has been a lot of policy work but little recent hard-core research. Of particular concern is financing by local governments and the institutions that allow cities to tax for current expenditures and borrow for capital projects such as infrastructure investments. As an example, in developing countries with weak institutions, metropolitan governments generally are not able to finance capital projects by borrowing either on bond markets or from international banks (given public infrastructure cannot be used as collateral). Borrowing is essential to efficient allocations given both limited current tax capacity and the fact that the benefits extend far into the future, so ideally financing is spread over time (Henderson and Venables, 2009). National governments can offer financing or guarantee loans, but then there is a problem of default by local governments on any loans granted to them. Of course, the national government can use grants to selectively finance local projects, but selection may be based on political considerations and less on local economic conditions. And national governments may be restricted in their revenue sources and ability to borrow as well. In short, it may be that many cities cannot access sufficient money and have deficient infrastructure investments (and some targeted cities may have excessive investments). We know of no hard core research on what the impact of underfunding (or overfunding) is on urban quality of life and growth of city populations or productivity. What are the productivity losses for a city such as Dar es Salaam with horrendous congestion, with little public transport, and with poor underfunded road networks? What will be the impact of the development of bus rapid transit now being constructed? We simply do not have findings from research which deals with such questions at a city or national scale.

In this section, we focus on two types of policies for which there is recent research and are fundamentally spatial in nature. The first policy concerns the causal effects of transport infrastructure investments linking cities and regions, as well as locations within cities, on urban form and city growth. The second policy concerns urban, or what we will label as big-city, bias in the allocation of public resources and the operation of markets.

22.4.3.1 Transport investments and technological change

Modeling suggests that transport infrastructure investments are responsible for changing patterns of specialization and growth of towns and regions observed in the data. This is the subject of a chapter in this handbook by [Redding and Turner \(2015\)](#), and our coverage is brief. An old debate concerns the effect on hinterland towns of improved linkages to the national centers of economic activity: linkages offer better access to markets but remove protection from outside competition for local producers. The work by [Donaldson \(2014\)](#) on historical India, which is based on the model of [Eaton and Kortum \(2002\)](#), suggests that transport investments lowering costs of trade between locations benefit essentially all cities or regions by allowing them to specialize in the production of goods for which they have more of a comparative advantage and to shed production of others and import them as transport costs fall.¹² For our purposes, the key is increased specialization on a wide-scale basis, consistent with the data on China and Korea we reviewed above. In the new economic geography models pioneered by [Krugman \(1991\)](#), transport improvements on a cruder scale lead first to centralization and specialization of the “core” region in manufacturing, consistent with the above analysis. But further improvements (at a later stage of development) can lead to decentralization of manufacturing to periphery regions if core regions become congested ([Puga, 1999](#)), as suggested by recent US data. In the simple new economic geography models, specialization and concentration are intertwined.

Transport investments also have strong effects on within-city decentralization of industry. In the United States, historically, goods moved across cities by rail, being shipped from terminals in or near city centers to other cities. Transport within the city to rail terminals by, for example, horse-drawn wagons was very expensive, so firms tended to cluster around the rail terminal in the city center. With the development of trucking and then the highway system, [Meyer et al. \(1965\)](#) argue that the construction of ring roads in cities in the 1950s and 1960s permitted various types of manufacturing to decentralize from urban cores to suburban areas with cheaper land and then to ship goods to rail sidings and suburban terminals by ring roads. For China, in a corresponding phase during the 1995–2008 period, [Baum-Snow et al. \(2013\)](#) show that rail and ring roads causally led to decentralization of manufacturing within Chinese cities.

22.4.3.2 Urban and political city bias

There is a development literature based on the two-sector model (e.g., see [Ray, 1998](#) for a synopsis) which talks about biases and/or policy distortions in labor or capital markets nationally that favor the urban sector and may draw in excessive numbers of migrants to

¹² Empirical work on China is less conclusive. [Faber \(2014\)](#) and [Banerjee et al. \(2012\)](#) reach opposite conclusions on the effect of transport improvements on the fortunes of hinterland areas that are “treated” with transport investments versus those that are not.

cities. Alternatively, there may be migration restrictions such as China's *hukou* system that restrain the extent of rural–urban migration. Here we turn to a related version of biases, where one city, or more generally, larger and politically connected cities are favored relative to other cities and the rural sector. As we will see, favoring a certain city may make that city either larger or smaller than it otherwise would be.

As reviewed by [Henderson \(1988\)](#) and [Duranton \(2008\)](#), the standard modeling of the effects of political bias assumes that favored cities are larger than they would be in the absence of favoritism. There is a system of cities in an economy of different types and equilibrium sizes. Under a developer regime, cities tend to operate near the peak of their inverted-U-shaped real income curves, at sizes where real incomes across different types of cities for a representative worker are equalized. In national labor markets, any one city faces a horizontal labor supply curve at that going real income. A city that is favored in capital markets or with special public services has an inverted-U shaped real income curve is shifted up—it can pay higher real income/utility at any size. If the city is subsidized for capital costs, that raises the marginal product of labor a competitive firm can pay. With unrestricted immigration, the size of the favored city expands beyond the peak of its inverted U. The equilibrium is the point down the right-hand side from the peak where city real income equals the going real income in national labor markets in other nonfavored cities. The implication is that, in a free-migration equilibrium, the benefits of favoritism are dissipated through increased commuting costs, or increased city disamenities more generally, as the city size expands beyond the peak potential real income point.

Empirically [Ades and Glaeser \(1995\)](#) and [Davis and Henderson \(2003\)](#) examine indirect evidence. As hinted at by the title of the article by Ades and Glaeser (trade and circuses), there seems to be a clear bias toward capital cities in many countries, especially before democratization. Relative to their economic position, they are much larger than other cities, indicating a bias toward investing in capital cities.

As a more specification example, for China there is indirect evidence given in [Au and Henderson \(2006b\)](#), who infer differential rates of return to the urban sector versus the rural sector and for different types of firms within the urban sector. Direct evidence is harder to find. While articles generate fiscal numbers showing higher *per capita* public expenditures in different classes of cities, it is hard to distinguish if that is bias, or if it is simply that it is efficient for public sectors to be larger in bigger cities, with their greater congestion and environmental issues. Capital markets where we expect an efficient allocation equalizes rates of return across cities can provide readier documentation. [Jefferson and Singh \(1999\)](#) estimate higher rates of return to rural-based firms compared with urban-based firms in the early 1990s in China. [Cai and Henderson \(2013\)](#) show that the rates of return to capital in China differ not only by firm type (lower for state-owned firms) but also by city type. All types of firms in political cities such as Beijing on average are favored (earn lower rates of return) than firms in ordinary-prefecture-level cities in

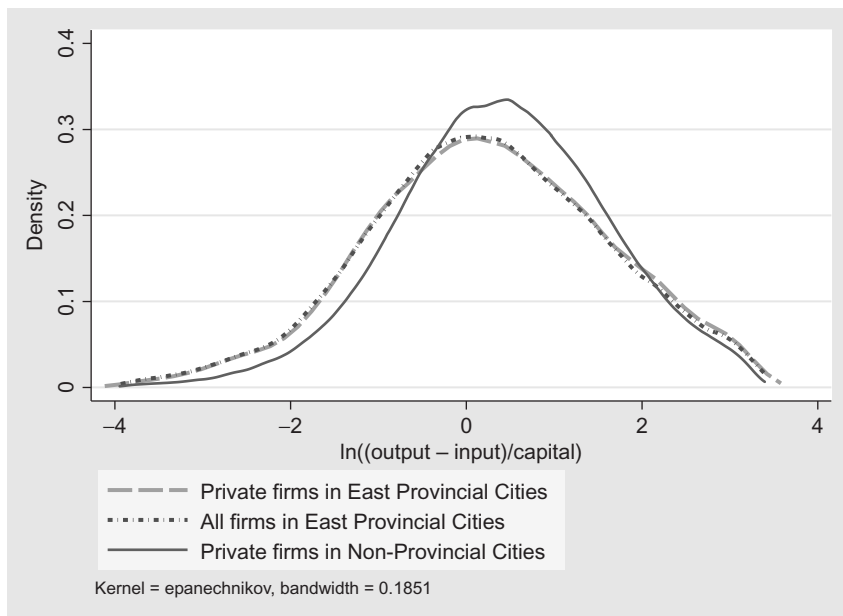


Figure 22.10 2007 Distribution of after-tax value added divided by net asset value (as proportional to the returns on capital).

China. [Figure 22.10](#) shows the distribution of returns for private sector firms in ordinary prefecture cities versus the three main provincial-level and heavily politically favored cities in eastern China, Beijing, Shanghai, and Tianjin.

Favoritism raises another critical issue. From the discussion of the inverted-U curve of real income against city size, it is apparent that cities would want to resist inward migration beyond the peak. If cities could price discriminate and city populations could be fixed, either “original” incumbent residents or a developer could restrict inward migration to the favored city and charge fees to marginal migrants ([Henderson, 1988](#); [Behrens and Robert-Nicoud, 2015](#)). According to the specific framework, city size is set to some real income maximizing size (for either the developer in a developer-controlled city or incumbent residents who control a city) between the peak and the free migration equilibrium. However, entry fees and price discrimination are not the direct institutions for cities in countries. Rather, it is through land markets and regulation that residents or city governments attempt to restrict city size.

In developed countries the tool to restrict size is exclusionary zoning. In the super-star cities article by Gyourko et al. (2013), favored cities, in their case cities favored with natural amenities, attempt to restrict inward migration through exclusionary zoning. Such zoning can effectively fix the number of dwelling units permitted in a locality. With that restriction, the key to entry is getting one of the fixed number of lots in the city. Lot prices rise so as to lower utility from entering the city to the outside option for the marginal

entrant. In their model, higher-skilled, higher-income people have a greater willingness to pay for the amenities of super-star cities. Thus, as the national population and real incomes rise, super-star cities both have higher price increases and a shift in the population composition toward higher-income people, who outbid others for the amenities of these cities. In these frameworks, the key assumption is that all dwelling units are provided in a formal sector governed by zoning laws.

In developing countries, the restrictions are different. Until recently in China, there have been explicit migration restrictions, directly limiting mobility. [Desmet and Rossi-Hansberg \(2013\)](#) find that the dispersion of amenities is greater across Chinese cities than across US cities, and show how this can be interpreted as evidence of migratory restrictions to some of the country's favored cities. In the absence of such restrictions, they find that some of China's largest cities would become even larger, and that overall welfare would increase significantly. Their article is also an example of how quantitative models in urban economics can be used to estimate the welfare impact of different policies.

Most countries do not have direct migration restrictions, and in China these are now disappearing. Restrictions take a different form. They involve land markets and the public sector, but not zoning, which restricts entry directly. In developing countries, there are informal housing sectors, which violate whatever regulations potentially govern the formal sector. As discussed in the chapter in this handbook by [Brueckner and Lall \(2015\)](#), governments in developing countries either do not have the power or political will to stop the development of informal sectors or permit them to develop in a second-best framework. Informal sectors may involve "squatting" ([Jimenez, 1984](#)), which means collective illegal seizure of land or illegal or quasi-legal development of land that is legally owned. One example of the latter is *loteamentos* in Brazil, which are developments in violation of national zoning laws but built on legally held land. Another example is the development of urban villages in cities in China. Urban villages are on land within the city that is still owned by a rural collective. Typically these were the former living areas of farm villages, where the city annexed the farm land but not the living area. These living areas are then intensively developed into high-density "slum" housing for migrants. This escape valve would then allow a free-migration equilibrium to emerge, but with one catch.

The catch involves the provision of local public services and becomes the basis for restricting inward migration. As [Cai \(2006\)](#) discusses for China, urban villages do not receive services from the city (central water or sewerage, garbage collection), and their children are generally excluded from state schools. This forces high-cost and/or low-quality provision of such services for migrants in these settlements, making migration much more costly for them. As nicely illustrated in [Duranton \(2008\)](#), effectively incumbent residents face one inverted U, while at the margin inward migrants face a different one that is shifted down. This reduces the population at which that national supply curve

of the population to the city intersects the effective real income curve for migrants. In China, this policy has been called “lifting the door sill” (Cai, 2006). For Brazil, Feler and Henderson (2011) attempt to estimate the causal effects on population growth especially of low-skilled people of denial of centralized water provision to likely migrant housing areas in Brazil. In Brazil, localities were not required to service areas which were not in the formal sector in the 1980s. This analysis suggests that the emergence of slum areas in cities in developing countries in some contexts reflects in part a strategic decision of localities to try to restrict inward migration, especially into favored larger cities such as national capitals (e.g., Beijing) or the seats of political elites (e.g., Shanghai and São Paulo).

22.5. CONCLUDING REMARKS

In this chapter, we have described recent theory and evidence of how the spatial distribution of economic activity changes as a country grows and develops. In doing so, we focused on different geographic units, starting with the coarse urban–rural distinction, then going to the entire distribution, and finishing with its upper tail—the cities. When addressing the question of how an economy’s spatial organization changes with development, the literature has often analyzed the long-term patterns of today’s developed countries, notably the United States. Undoubtedly, the past spatial development of the United States holds valuable lessons for today’s developing countries, so this strategy is often both useful and appropriate. At the same time, today’s world is different from the one faced by the United States and other developed countries in the nineteenth and twentieth centuries. For example, the increasing impact of trade may imply that some countries can urbanize without industrializing. Traditionally the paucity of geographically disaggregated data has limited the extent of empirical analysis on developing countries. However, the rapidly increasing availability of data, together with geographic information system tools, is changing this. In fact, as this chapter has made clear, the last decade has seen a growing number of empirical studies using data from developing countries. We believe that there is a need for more such studies in order to elicit the stylized facts which should form the basis for further theoretical work on the link between geography and development. More work is also needed in modeling and understanding the relationship between space and development. The spatial distribution of economic activity affects growth, and vice versa. An economy’s degree of urbanization is not only a consequence of its development, it is also a determinant of its growth. To understand better these links, there is a need for more micro studies and for more quantitative work with an emphasis on counterfactual policy experiments. It is clear that a country’s spatial organization is not independent of its macroeconomic performance. Regional and urban economists should therefore continue their efforts to develop the tools needed to inform policy makers of how regional and spatial policies affect welfare and growth.

REFERENCES

- Aarland, K.J., Davis, J.C., Henderson, J.V., Ono, Y., 2007. Spatial organization of firms. *Rand J. Econ.* 38, 480–494.
- Abdel-Rahman, H., Anas, A., 2004. Theories of systems of cities. In: Henderson, J.V., Thisse, J.F. (Eds.), *Handbook of Regional and Urban Economics*, vol. 4. Elsevier, Amsterdam, pp. 2293–2339.
- Abdel-Rahman, H., Fujita, M., 1993. Specialization and diversification in a system of cities. *J. Urban Econ.* 33, 189–222.
- Ades, A.F., Glaeser, E.L., 1995. Trade and circuses: explaining urban giants. *Q. J. Econ.* 110, 195–227.
- Allen, R.C., 2004. Agriculture during the industrial revolution, 1700–1850. In: Floud, R., Johnson, P. (Eds.), *The Cambridge Economic History of Modern Britain*, vol. 1. Cambridge University Press, Cambridge, United Kingdom. **Industrialisation 1700–1860 (Chapter 1)**.
- Arthur, B., 1990. Silicon valley locational clusters: when do increasing returns to scale imply monopoly. *Math. Soc. Sci.* 19, 235–251.
- Au, C.C., Henderson, J.V., 2006. Are Chinese cities too small? *Rev. Econ. Stud.* 73, 549–576.
- Au, C.C., Henderson, J.V., 2006. How migration restrictions limit agglomeration and productivity in China. *J. Econ. Dev.* 80, 350–388.
- Bairoch, P., Batou, J., Chèvre, P., 1988. *La population des villes européennes de 800 à 1850*. Centre d'Histoire Economique Internationale de l'Université de Genève, Librairie Droz.
- Baldwin, R.E., Martin, P., 2004. Agglomeration and regional growth. In: Henderson, J.V., Thisse, J.F. (Eds.), *Handbook of Regional and Urban Economics*. first ed., vol. 4. Elsevier, Amsterdam, pp. 2671–2711 (Chapter 60).
- Banerjee, A., Duflo, E., Qian, N., 2012. On the road: transportation infrastructure and economic development. NBER Working paper 17897.
- Basu, S., Fernald, J., 2007. Information and communications technology as a general purpose technology: evidence from US industry data. *Ger. Econ. Rev.* 8 (2), 146–173.
- Baum-Snow, N., Pavin, R., 2012. Understanding the city size wage gap. *Rev. Econ. Stud.* 79, 88–127.
- Baum-Snow, N., Brandt, L., Henderson, J.V., Turner, M., Zhang, Q., 2013. Roads, railways and decentralization of Chinese cities. Brown University, processed.
- Becker, R., Henderson, J.V., 2000. Intra-industry specialization and urban development. In: Huriot, J.M., Thisse, J. (Eds.), *The Economics of Cities: Theoretical Perspectives*. Cambridge University Press, Cambridge, UK, pp. 138–166.
- Beeson, P.E., DeJong, D.N., 2002. Divergence. *Contrib. Macroecon.* 2 (1), Article 6, B.E. Press.
- Behrens, K., Robert-Nicoud, F., 2015. Agglomeration theory with heterogeneous agents. In: Duranton, G., Henderson, J.V., Strange, W. (Eds.), *Handbook of Regional and Urban Economics*, vol. 5. Elsevier, Amsterdam.
- Behrens, K., Duranton, G., Robert-Nicoud, F., 2014. Productive cities: sorting, selection, and agglomeration. *J. Polit. Econ.* 122, 507–553.
- Black, D., Henderson, J.V., 1999. A theory of urban growth. *J. Polit. Econ.* 107 (2), 252–284.
- Black, D., Henderson, J.V., 1999. Spatial evolution of population and industry in the USA. *Am. Econ. Rev. Pap. Proc.* 89 (2), 321–327.
- Black, D., Henderson, J.V., 2003. Urban evolution in the USA. *J. Econ. Geogr.* 3, 343–372.
- Boucekkine, R., Camacho, C., Zou, B., 2009. Bridging the gap between growth theory and the new economic geography: the spatial Ramsey model. *Macroecon. Dyn.* 13, 20–45.
- Brock, W., Xepapadeas, A., 2008. Diffusion-induced instability and pattern formation in infinite horizon recursive optimal control. *J. Econ. Dyn. Control.* 32, 2745–2787.
- Brock, W., Xepapadeas, A., 2010. Pattern formation, spatial externalities and regulation in coupled economic-ecological systems. *J. Environ. Econ. Manag.* 59, 149–164.
- Brueckner, J., Lall, S., 2015. Cities in developing countries: fueled by rural-urban migration, lacking in tenure security, and short of affordable housing. In: Duranton, G., Henderson, J.V., Strange, W. (Eds.), *Handbook of Regional and Urban Economics*, vol. 5. Elsevier, Amsterdam.
- Brühlhart, M., Sbergami, F., 2009. Agglomeration and growth: cross-country evidence. *J. Urban Econ.* 65 (1), 48–63.
- Cai, F., 2006. Floating populations: urbanization with Chinese characteristics. CASS mimeo.

- Cai, W., Henderson, J.V., 2013. The Bias towards political cities and state owned firms in China's capital markets. LSE, processed.
- Carlino, G., Kerr, W., 2015. Agglomeration and innovation. In: Duranton, G., Henderson, J.V., Strange, W. (Eds.), *Handbook of Regional and Urban Economics*, vol. 5. Elsevier, Amsterdam.
- Caselli, F., Coleman II., W.J., 2001. The U.S. structural transformation and regional convergence: a reinterpretation. *J. Polit. Econ.* 109, 584–616.
- Caselli, P., Paternò, F., 2001. ICT accumulation and productivity growth in the United States: an analysis based on industry data. *Temi di Discussione* 419, Banco d'Italia.
- Chun, H., Kim, J.W., Lee, J., Morck, R., 2005. Information technology, creative destruction, and firm-specific volatility. Unpublished manuscript.
- Costinot, A., 1999. An elementary theory of comparative advantage. *Econometrica* 77, 1165–1192.
- David, P.A., Wright, G., 2003. General purpose technologies and surges in productivity: historical reflections on the future of the ICT revolution. In: David, P.A., Thomas, M. (Eds.), *The Economic Future in Historical Perspective*. Oxford University Press, Oxford, UK.
- Davis, D., Dingel, J., 2013. The comparative advantage of cities. Columbia University, processed.
- Davis, S., Haltiwanger, J., 1998. Measuring gross worker and job flows. In: Haltiwanger, J.C., Manser, M.E., Topele, R.H. (Eds.), *Labor Statistics Measurement Issues*. University of Chicago Press, Chicago.
- Davis, J., Henderson, J.V., 2003. Evidence on the political economy of the urbanization process. *J. Urban Econ.* 53, 98–125.
- Davis, J., Henderson, J.V., 2008. Agglomeration of headquarters. *Reg. Sci. Urban Econ.* 63, 431–450.
- Desmet, K., Fafchamps, M., 2005. Changes in the spatial concentration of employment across U.S. counties: a sectoral analysis 1972–2000. *J. Econ. Geogr.* 5, 261–284.
- Desmet, K., Fafchamps, M., 2006. Employment concentration across U.S. counties. *Regional Sci. Urban Econ.* 36, 482–509.
- Desmet, K., Parente, S.L., 2012. The evolution of markets and the revolution of industry: a unified theory of growth. *J. Econ. Growth* 17, 205–234.
- Desmet, K., Rappaport, J., 2013. The settlement of the United States, 1800–2000: the long transition to Gibrat's law. *CEPR Discussion Paper #9353*.
- Desmet, K., Rossi-Hansberg, E., 2009. Spatial growth and industry age. *J. Econ. Theory* 144, 2477–2502.
- Desmet, K., Rossi-Hansberg, E., 2010. On spatial dynamics. *J. Reg. Sci.* 50, 43–63.
- Desmet, K., Rossi-Hansberg, E., 2012. Innovation in space. *Am. Econ. Rev. Pap. Proc.* 102, 447–452.
- Desmet, K., Rossi-Hansberg, E., 2013. Urban accounting and welfare. *Am. Econ. Rev.* 103, 2296–2327.
- Desmet, K., Rossi-Hansberg, E., 2014a. Spatial development. *Am. Econ. Rev.* 104, 1211–1243.
- Desmet, K., Rossi-Hansberg, E., 2014b. On the spatial economic impact of global warming. Working paper.
- Desmet, K., Ghani, E., O'Connell, S., Rossi-Hansberg, E., 2015. The spatial development of India. *J. Reg. Sci.* 55, 10–30.
- Diamond, J., 1997. *Guns, Germs, and Steel: The Fates of Human Societies*. W.W. Norton, New York.
- Doepke, M., 2004. Accounting for fertility decline during the transition to growth. *J. Econ. Growth* 9, 347–383.
- Donaldson, D., 2014. Railroads of the raj: estimating the impact of transportation infrastructure. *Am. Econ. Rev.*, forthcoming.
- Duranton, G., 2007. Urban evolutions: the fast, the slow, and the still. *Am. Econ. Rev.* 97, 197–221.
- Duranton, G., 2008. Viewpoint: from cities to productivity and growth in developing countries. *Can. J. Econ.* 41, 689–736.
- Duranton, G., Overman, H.G., 2005. Testing for localization using micro-geographic data. *Rev. Econ. Stud.* 72, 1077–1106.
- Duranton, G., Overman, H.G., 2008. Exploring the detailed location patterns of U.K. manufacturing industries using microgeographic data. *J. Reg. Sci.* 48, 213–243.
- Duranton, G., Puga, D., 2001. Nursery cities. *Am. Econ. Rev.* 91, 1454–1477.
- Duranton, G., Puga, D., 2004. Micro-foundations of urban agglomeration economies. In: Henderson, J.V., Thisse, J.F. (Eds.), *Handbook of Regional and Urban Economics*, vol. 4. Elsevier, Amsterdam, pp. 2063–2117.
- Duranton, G., Puga, D., 2005. From sectoral to functional urban specialisation. *J. Urban Econ.* 57, 343–370.
- Duranton, G., Puga, D., 2014. The growth of cities. In: Durlauf, S.N., Aghion, P. (Eds.), *Handbook of Economic Growth*, vol. 2. Elsevier, Amsterdam.

- Duranton, G., Turner, M., 2012. Urban growth and transportation. *Rev. Econ. Stud.* 79, 1407–1440.
- Easterlin, R.A., 1960. Interregional difference in per capita income, population, and total income, 1840–1950. In: Parker, W. (Ed.), *Trends in the American Economy in the Nineteenth Century*, Studies in Income and Wealth. Princeton University Press, vol. 24. Princeton, NJ, pp. 73–140.
- Eaton, J., Eckstein, Z., 1997. Cities and growth: evidence from France and Japan. *Reg. Sci. Urban Econ.* 27, 443–474.
- Eaton, J., Kortum, S., 2002. Technology, geography, and trade. *Econometrica* 70, 1741–1779.
- Eeckhout, J., 2004. Gibrat's law for (all) cities. *Am. Econ. Rev.* 94, 1429–1451.
- Eeckhout, J., Pinheiro, R., Schmidheiny, K., 2014. Spatial sorting. *J. Polit. Econ.* 122, 554–620.
- Ellison, G., Glaeser, E.L., 1997. Geographic concentration in U.S. manufacturing industries: a dartboard approach. *J. Polit. Econ.* 105, 889–927.
- Faber, B., 2014. Trade integration, market size, and industrialization: evidence from China's National Trunk Highway System. *Rev. Econ. Stud.* forthcoming.
- Fafchamps, M., Shilpi, F., 2005. Cities and specialization: evidence from South Asia. *Econ. J.* 115, 477–504.
- Fallah, B., Partridge, M., 2012. Geography and high-tech employment growth in U.S. counties. MPRA Paper 38294.
- Fay, M., Opal, C., 2000. Urbanization without growth: a not-so-uncommon phenomenon. World Bank Policy Research Working paper Series 2412.
- Feler, L., Henderson, J.V., 2011. Exclusionary policies in urban development. *J. Urban Econ.* 69, 253–272.
- Forman, C., Goldfarb, A., Greenstein, S., 2005. Geographic location and the diffusion of internet technology. *Electron. Commer. Res. Appl.* 4, 1–13.
- Fujita, M., Ogawa, H., 1982. Multiple equilibria and structural transition of non-monocentric configurations. *Reg. Sci. Urban Econ.* 12, 161–196.
- Fujita, M., Henderson, J.V., Kanemoto, Y., Mori, T., 2004. The spatial distribution of economic activities in Japan and China. In: Henderson, J.V., Thisse, J.F. (Eds.), *Handbook of Regional and Urban Economics*, vol. 4. Elsevier, Amsterdam.
- Gabaix, X., 1999. Zipf's law for cities: an explanation. *Q. J. Econ.* 114, 739–767.
- Gabaix, X., Ioannides, Y., 2004. The evolution of city size distributions. In: Henderson, J.V., Thisse, J.F. (Eds.), *Handbook of Regional and Urban Economics*, vol. 4. Elsevier, Amsterdam.
- Galor, O., Weil, D., 2000. Population, technology, and growth: from the Malthusian regime to the demographic transition and beyond. *Am. Econ. Rev.* 90, 806–828.
- Galor, O., Moav, O., Vollrath, D., 2009. Inequality in landownership, human capital promoting institutions and the great divergence. *Rev. Econ. Stud.* 76 (1), 143–179.
- Gibrat, R., 1931. *Les inégalités économiques: applications aux inégalités de richesses, à la concentration des entreprises, aux populations des villes, aux statistiques des familles, etc., d'une loi nouvelle, la loi de l'effet proportionnel*. Librairie du Recueil Sirey, Paris.
- Glaeser, E., Gyourko, J., 2005. Urban decline and durable housing. *J. Polit. Econ.* 113, 345–375.
- Gollin, D., Parente, S.L., Rogerson, R., 2007. The food problem and the evolution of international income levels. *J. Monet. Econ.* 54, 1230–1255.
- Gyourko, J., Mayer, C., Sinai, T., 2013. Superstar cities. *Am. Econ. J. Econ. Policy* 5, 167–199.
- Hansen, G., Prescott, E.C., 2002. Malthus to Solow. *Am. Econ. Rev.* 92, 1205–1217.
- Harris, J.R., Todaro, M.P., 1970. Migration, unemployment and development: a two-sector analysis. *Am. Econ. Rev.* 60, 126–142.
- Harris-Dobkins, L., Ioannides, Y.M., 2001. Spatial interactions among U.S. cities: 1900–1990. *Reg. Sci. Urban Econ.* 31, 701–731.
- Helsley, R.W., Strange, W.C., 1990. Agglomeration economies and matching in a system of cities. *Reg. Sci. Urban Econ.* 20, 189–212.
- Helsley, R.W., Strange, W.C., 2014. Coagglomeration, clusters and the scale and composition of cities. *J. Polit. Econ.* 122, 1064–1093.
- Henderson, J.V., 1974. The sizes and types of cities. *Am. Econ. Rev.* 64 (4), 640–656.
- Henderson, J.V., 1988. *Urban Development: Theory, Fact, and Illusion*. Oxford University Press, New York.
- Henderson, J.V., 1997. Medium size cities. *Reg. Sci. Urban Econ.* 27, 583–612.

- Henderson, J.V., 2005. Urbanization and growth. In: Aghion, P., Durlauf, S. (Eds.), *Handbook of Economic Growth*. Elsevier, Amsterdam.
- Henderson, J.V., 2010. Cities and development. *J. Reg. Sci.* 50, 515–540.
- Henderson, J.V., Venables, A., 2009. The dynamics of city formation. *Rev. Econ. Dyn.* 12, 233–254.
- Henderson, J.V., Wang, H.G., 2005. Aspects of the rural–urban transformation of countries. *J. Econ. Geogr.* 5, 23–42.
- Henderson, J.V., Wang, H.G., 2007. Urbanization and city growth: the role of institutions. *Reg. Sci. Urban Econ.* 37, 283–313.
- Henderson, J.V., Lee, T., Lee, Y.J., 2001. Scale externalities in a developing country. *J. Urban Econ.* 49, 479–504.
- Henderson, J.V., Roberts, M., Storeygard, A., 2013. Is urbanization in sub-Saharan Africa different? Policy Research Working paper Series 6481. World Bank.
- Hobijn, B., Jovanovic, B., 2001. The information–technology revolution and the stock market: evidence. *Am. Econ. Rev.* 91, 1203–1220.
- Holmes, T.J., Lee, S., 2010. Cities as six-by-six-mile squares: Zipf's law? NBER. In: Glaeser, E. (Ed.), *Agglomeration Economics*. National Bureau of Economic Research, pp. 105–131 (Chapter).
- Holmes, T., Stevens, J.J., 2004. Spatial distribution of economic activities in North America. In: Henderson, J.V., Thisse, J.F. (Eds.), *Handbook of Regional and Urban Economics*, vol. 4. Elsevier, Amsterdam.
- Ioannides, Y., Skouras, S., 2013. US city size distribution: robustly Pareto, but only in the tail. *J. Urban Econ.* 73 (1), 18–29.
- Jefferson, G.H., Singh, I., 1999. *Enterprise Reform in China: Ownership, Transition, and Performance*. World Bank, Washington, DC and Oxford University Press, New York.
- Jimenez, E., 1984. Tenure security and urban squatting. *Rev. Econ. Stat.* 66, 556–567.
- Jovanovic, B., Rousseau, P.L., 2005. General purpose technologies. In: Aghion, P., Durlauf, S. (Eds.), *Handbook of Economic Growth*. Elsevier, Amsterdam (Chapter 18).
- Kim, S., 1998. Economic integration and convergence: U.S. regions, 1840–1987. *J. Econ. Hist.* 58, 659–683.
- Kim, S., 1999. Regions, resources, and economic geography: the sources of USA comparative advantage, 1880–1987. *Reg. Sci. Urban Econ.* 29, 1–32.
- Kim, S., 2009. Spatial inequality and development: theories, facts and policies. In: Buckley, R., Annez, P., Spence, M. (Eds.), *Urbanization and Growth*. The International Bank for Reconstruction and Development and World Bank, Washington, DC.
- Kim, S., Margo, R.A., 2004. Historical perspectives on U.S. economic geography. In: Henderson, J.V., Thisse, J.F. (Eds.), *Handbook of Regional and Urban Economics*. first ed., vol. 4. Elsevier, Amsterdam, pp. 2981–3019 (Chapter 66).
- Kolko, J., 1999. Can I get some service here: information technology, service industries, and the future of cities. Harvard University, Mimeo.
- Krugman, P., 1991. Increasing returns and economic geography. *J. Polit. Econ.* 99 (3), 483–499.
- Krugman, P., 1992. *Geography and Trade*. MIT Press, Gaston Eyskens Lecture Series, Cambridge, MA.
- Krugman, P., 1996. Confronting the mystery of urban hierarchy. *J. Jpn. Int. Econ.* 10, 399–418.
- Lee, S., Li, Q., 2013. Uneven landscapes and city size distributions. *J. Urban Econ.* 78, 19–29.
- Lewis, W.A., 1954. Economic development with unlimited supplies of labour. *Manch. Sch.* 22 (2), 139–191.
- Lucas, R.E., 2004. Life earnings and rural–urban migration. *J. Polit. Econ.* 112 (S1), S29–S59.
- Matsuyama, K., 1992. Agricultural productivity, comparative advantage, and economic growth. *J. Econ. Theory* 58, 317–334.
- McGuckin, R.H., Stiroh, K.J., 2002. Computers and productivity: are aggregation effects important? *Econ. Inq.* 40, 42–59.
- Meyer, J.R., Kain, J.F., Wohl, M., 1965. *The Urban Transportation Problem*. Harvard University Press, Cambridge.
- Michaels, G., Rauch, F., Redding, S., 2012. Urbanization and structural transformation. *Q. J. Econ.* 127, 535–586.

- Mitchener, K.J., McLean, I.W., 1999. U.S. regional growth and convergence, 1880–1980. *J. Econ. Hist.* 59, 1016–1042.
- Ngai, R.L., Pissarides, C.A., 2007. Structural change in a multisector model of growth. *Am. Econ. Rev.* 97 (1), 429–443.
- Nurkse, R., 1953. *Problems of Capital Formation in Underdeveloped Countries*. Oxford University Press, New York.
- Ono, Y., 2003. Outsourcing business services and the role of central administrative offices. *J. Urban Econ.* 53 (3), 377–395.
- Puga, D., 1999. The rise and fall of regional inequalities. *Eur. Econ. Rev.* 43, 303–334.
- Rappaport, J., Sachs, J.D., 2003. The United States as a coastal nation. *J. Econ. Growth* 8, 5–46.
- Rauch, J.E., 1993. Does history matter only when it matters a little? The case of city–industry location. *Q. J. Econ.* 108, 843–867.
- Ray, D., 1998. *Development Economics*. Princeton University Press, Princeton (Chapter 3).
- Redding, S., Turner, M., 2015. Transportation costs and the spatial organization of economic activity. In: Duranton, G.J., Henderson, V., Strange, W. (Eds.), *Handbook of Regional and Urban Economics*, vol. 5. Elsevier, Amsterdam.
- Renaud, B., 1981. *National Urbanization Policy in Developing Countries*. Oxford University Press, Oxford.
- Rosenbloom, J.L., 1990. One market or many? Labor market integration in the late nineteenth-century United States. *J. Econ. Hist.* 50, 85–107.
- Rossi-Hansberg, E., Wright, M., 2007. Urban structure and growth. *Rev. Econ. Stud.* 74 (2), 597–624.
- Rostow, W.W., 1960. *The Stages of Economic Growth: A Non-Communist Manifesto*. Cambridge University Press, Cambridge, UK.
- Schultz, T.W., 1968. *Economic Growth and Agriculture*. McGraw-Hill, New York.
- Shaw-Taylor, L., Wrigley, E.A., 2008. The Occupational Structure of England c.1750 to c.1871, <http://www.geog.cam.ac.uk/research/projects/occupations/>.
- Tabuchi, T., Thisse, J.F., 2011. A new economic geography model of central places. *J. Urban Econ.* 69, 240–252.
- Tamura, R., 2002. Human capital and the switch from agriculture to industry. *J. Econ. Dyn. Control.* 27, 207–242.
- Tecu, I., 2013. The location of industrial innovation: does manufacturing matter? US Census Bureau Center for Economic Studies Paper No. CES-WP-13-09.
- Trew, A., 2014. Spatial takeoff in the first industrial revolution. *Rev. Econ. Dyn.* 17, 707–725.
- Triplett, J.E., Bosworth, B.P., 2002. ‘Baumol’s Disease’ has been cured: IT and multifactor productivity in U.S. services industries. In: Jansen, D.W. (Ed.), *The New Economy and Beyond: Past, Present, and Future*. Edward Elgar Publishing, Cheltenham, UK and Northampton, MA, pp. 34–71.
- United Nations, 2010. *World Population Policies 2009*. United Nations, New York.
- Williamson, J., 1965. Antebellum urbanization in the American Northeast. *J. Econ. Hist.* 25, 592–608.
- World Bank, 2009. *World Development Report 2009: Reshaping Economic Geography*. The World Bank, Washington, DC.

This page intentionally left blank

CHAPTER 23

Urban Crime

Brendan O'Flaherty^{*}, Rajiv Sethi^{†,‡}

^{*}Department of Economics, Columbia University, NY, USA

[†]Department of Economics, Barnard College, Columbia University, NY, USA

[‡]Santa Fe Institute, Santa Fe, NM, USA

Contents

23.1. Introduction	1521
23.2. Criminogenic Characteristics	1522
23.2.1 The strong claim of criminogenic characteristics	1525
23.2.1.1 <i>Age and gender</i>	1525
23.2.1.2 <i>Childhood lead exposure</i>	1525
23.2.1.3 <i>Schizophrenia, ADHD, and major depressive disorder</i>	1526
23.2.1.4 <i>Family structure</i>	1526
23.2.1.5 <i>Abortion</i>	1527
23.2.2 The weak claim of criminogenic characteristics	1528
23.2.2.1 <i>Education</i>	1529
23.2.2.2 <i>Psychological and character traits</i>	1531
23.2.2.3 <i>Brain functioning, brain structure, in utero experience, and genetics</i>	1532
23.2.2.4 <i>Other characteristics</i>	1534
23.2.3 Race and ethnicity	1534
23.3. Incentives and Deterrence	1536
23.3.1 Certainty and severity	1537
23.3.2 Empirical studies of deterrence	1539
23.3.2.1 <i>Certainty studies</i>	1540
23.3.2.2 <i>Severity studies</i>	1545
23.3.2.3 <i>Violence reduction initiatives</i>	1546
23.3.2.4 <i>What empirical studies tell us</i>	1547
23.3.3 Income, earnings, and employment	1547
23.3.4 Race and incentives	1550
23.4. Interactions	1552
23.4.1 Private actions and displacement	1553
23.4.2 Deterrence and selection	1555
23.4.3 Stereotypes	1556
23.4.4 Segregation	1557
23.4.5 Peer effects and strategic complementarity	1559
23.4.6 The castle doctrine and stand-your-ground laws	1562
23.4.7 Police stops	1562
23.5. Incarceration	1567
23.5.1 Accounting for trends	1568
23.5.2 Empirical evidence on incapacitation	1571
23.5.2.1 <i>The weak claim of incapacitation</i>	1572

23.5.2.2	<i>The strong claim of incapacitation</i>	1572
23.5.2.3	<i>Incapacitation plus deterrence effects</i>	1574
23.5.3	Postprison effects	1575
23.5.4	Prison gangs	1578
23.5.4.1	<i>Preprison effects and prison culture</i>	1579
23.5.5	Trade-offs	1580
23.5.6	Race and mass incarceration	1581
23.6.	Big Swings in Crime	1583
23.6.1	Stylized facts	1584
23.6.2	Explaining the American crime drop of the 1990s	1586
23.6.2.1	<i>Traditional explanations</i>	1587
23.6.2.2	<i>Nontraditional explanations</i>	1589
23.6.2.3	<i>Wrapping up: the great American crime decline of 1991–2001</i>	1599
23.6.3	The modest American crime decline of the twenty-first century	1599
23.6.4	Developed countries in the twenty-first century	1602
23.6.5	The great American crime rise, 1965–1975	1602
23.7.	Where are Crimes Committed?	1604
23.7.1	Intermetropolitan variation	1604
23.7.2	Intrametropolitan variation	1606
23.7.3	Explaining intrametropolitan variation	1608
23.7.4	Interjurisdiction variation	1611
23.8.	Conclusions	1612
	Acknowledgments	1613
	References	1613

Abstract

We survey the literature on index crime, paying particular attention to spatial issues. We note the contrasting descriptive traditions of Lombroso (characteristics matter) and Beccaria (incentives matter), and the contrasting policy traditions of incapacitation (predict who will offend and keep them from doing it) and deterrence (uncover who offended and punish them). The economics of crime has several points of contact with the economics of space, since the commission of an index crime requires proximity between offenders and victims (or their property). We explore these linkages, as well as a range of other issues: the effects of certainty and severity of punishment on crime; the role of stereotypes in interactions between offenders, victims, and law enforcement officers; and racial disparities in victimization, offending, and incarceration. The economics of crime has made tremendous progress, but enormous variation across both time and space remains poorly understood, and many nontraditional explanations often neglected by economists need to be explored more systematically.

Keywords

Crime, Victimization, Offending, Monitoring, Enforcement

JEL Classification Codes

K42, H76, R12

23.1. INTRODUCTION

Crimes are activities that governments have threatened to punish rather severely. This threat, in principle, is not contingent on whether any third party complains about the activity. If there is no government, or no prospect of punishment, there is no crime.

Across time and space, the kinds of activities that are crimes vary widely. Sodomy and blasphemy were once crimes in New York, and they are still crimes in many parts of the world. Married men in New York could once inflict violence on their wives with impunity, and continue to have that latitude elsewhere. Some activities that have modest external costs are crimes in most American cities—trading safe sex for money, or stealing a carton of cigarettes—while other activities with considerable external costs are not—sticking an umbrella in the closing door of a rush-hour subway train, running an unsafe fertilizer plant in Texas, or driving while sober and absent-mindedly killing a pedestrian, for instance.

Many activities that almost all governments proscribe are extremely noxious to society, and cities probably could not flourish unless they were somehow restrained. But threatening punishments credibly and carrying them out are also expensive, both for the government and for the people punished and their families. These two costs—those of the activities that are proscribed, and those of the punishments that make them crimes—are central to the economic analysis of crime.

Urban economists have focused on the so-called index crimes and so will this review. These are the acts in the FBI's index of crime—murder, rape, robbery, assault, larceny, burglary, and motor vehicle theft.¹ These crimes require that the offender be near the victim or the victim's property, and so have an inherently spatial component. For the most part, they are also *mala in se* and proscribed in almost all societies. The boundaries of these crimes, however, are not always well established, for instance in cases involving vehicular homicide, justifiable homicide, euthanasia, date rape, and marital rape. We will look at crimes involving illicit drugs too, since they are tied to the index crimes in many ways, both real and imaginary.

Our focus on index crimes is driven mainly by convenience, and by the way that law enforcement agencies are organized. But because these are probably the most serious crimes with a strong spatial component, the focus is probably not misplaced. White-collar crimes such as fraud and embezzlement transfer vastly more resources (Anderson, 1999), and cybercrimes are a rapidly growing problem, but where you live, whom you live with, and whether a police officer is standing outside your home probably exert little influence over whether you fall victim to these crimes. We will also largely omit terrorism. Terrorists commit index crimes, but their preferences and the strategies that might deter

¹ Following an act of Congress, arson has legally been considered an index crime since 1979, but the numbers are small and detection uncertain. We will disregard arson in this chapter.

them are different enough from those of more prosaic criminals that they deserve study in a separate place.²

If index crimes are committed, they have to be committed somewhere, and so both strategies for committing them and strategies for avoiding them must take location into account. Ultimately one of our main concerns is what makes particular times and places more dangerous than others, and [Section 23.6](#) (on times) and [Section 23.7](#) (on places) are explicitly devoted to this question. As we show there, the economic literature on the timing and the place of crime is scant, with many important issues remaining unresolved. Most of this review can be seen as looking at various hypotheses for spatiotemporal disparities, even though the solid work on these hypotheses rarely develops far enough to say much about time and place. Perhaps locations are dangerous because of the reasonably stable characteristics of the people who live or work close to them; thus, [Section 23.2](#) examines what we know about the relationship between criminogenic characteristics and crime. Perhaps locations are dangerous because of poor incentives nearby, either negative (not enough jobs) or positive (not enough cops); [Sections 23.3](#) and [23.5](#) look at how incentives effect crime. Perhaps more complex interactions and the behavior of potential victims create danger, and [Section 23.4](#) accordingly studies interaction structures.

In this review, we concentrate on the incidence of crime and the effectiveness of efforts to reduce it. We do not provide a systematic discussion of the welfare implications of crime and crime control policies. This is because of both space constraints and the philosophical difficulty that is attached to the issue. How, for instance, do we assess the benefits that an offender derives from crime? If fear of punishment deters a would-be assailant, is his/her welfare loss a social cost? If it is not, then should the charges incurred by power companies compelled to reduce emissions also be excluded from any analysis of welfare? [Coase \(1960\)](#) famously observed that harmful actions have a reciprocal character—prevention of harm to one party by another is itself harmful to the party that is restrained. Thus, the costs of abatement incurred by a polluting firm are generally treated as social costs, to be fully accounted for in any welfare analysis. But this principle is seldom invoked in the welfare analysis of criminal behavior. Such issues deserve careful scrutiny, but lie beyond the scope of the more modest exercise undertaken here.

23.2. CRIMINOGENIC CHARACTERISTICS

The common-sense, man-in-the-street reaction to crime is to observe that most of us do not go around murdering our spouses or molesting small children, and to ask what makes criminals different. If criminals commit crimes because of their individual characteristics, then the volume of crime is roughly proportional to the number of criminals. Crime goes

² Unexpected terrorist attacks can lead to sudden and plausibly exogenous changes in police deployments, and this can be used to identify causal effects of police presence on index crime. We discuss this literature extensively below, but avoid a discussion of the causes and consequences of terrorism itself.

up when the number of criminals goes up, and goes down when the number of criminals goes down. Crime is high in places with a lot of criminals and is low in places with few criminals. From this perspective, the objective of research should be to determine what makes criminals, and the objective of policy should be to reduce the number of people with the characteristics that make them criminals, or at least to make sure they are not in contact with potential victims.

This view of crime has a long history. Cesare Lombroso (1835–1909), who is often called the father of modern criminology, maintained that criminals were “atavistic,” born different from other people, and could be distinguished by their physical features, many of them simian, such as a sloping forehead, ears of unusual size, facial asymmetry, left-handedness, prognathism, and excessive arm length. Lombroso’s views have largely been discredited, but the notion remains pervasive that some relatively permanent features, either genetic or acquired, predispose people to commit crimes.³

It is probably more useful to us to know whether the large stranger approaching us on a quiet street is likely to rob us than it is to know why car thefts declined in Chicago in the 1990s; thus the practical appeal for the general population of this line of thinking is immediate. The facts that only a small proportion of the population commits index crimes in any year and that the committing of crime is serially correlated suggest that something about people who commit crimes sets them apart from others and makes it appropriate to talk about “criminals” rather than simply about “people who commit crimes.”⁴

This view of crime has policy implications. If the characteristics that incline people to commit crimes are known, then crime can be prevented either by eliminating those characteristics or by separating the people with those characteristics from potential victims. Prediction leads to prevention through incapacitation.

Government policies that are derived from this premise, however, face some thorny ethical issues. Some of these policies harm the people that are likely to commit crimes—sending them to prison, for instance, or stopping them on the street for humiliating searches. Harming innocent people to prevent crimes they have not committed strikes many—including the framers of the US Bill of Rights—as an activity governments have no right to be involved in. Policies that help people who are highly likely to commit crimes—providing more recreational opportunities, for instance, for boys than girls (since women commit far fewer crimes than men do)—run into the same difficulty. Should governments deny people resources because they lack criminogenic characteristics?

³ This notion continues to stimulate research. For instance, [Valla et al. \(2011\)](#) report that twenty-first century college students can do significantly better than chance in identifying people who have been convicted of crimes from anonymous head shots of white male 20–25-year-olds.

⁴ Of all people who were 18 years old in California in 1974, only 24% were arrested at least once over the next 11 years—and this is the age group with the highest propensity to commit crime in an era of high crime. However, contingent on being arrested, the probability of being arrested at least one more time was 45.9%. The corresponding figures for index crimes were 10.5% and 36.7%, respectively ([Tillman, 1987](#)).

In this section, we examine views of crime that emphasize relatively stable characteristics of criminals. In the first part, we look at what we call the *strong claim of criminogenic characteristics*, that the number of people with certain traits affects the volume of crime. Relatively few traits have been shown to cause crime in this sense: being a young man, growing up breathing atmospheric lead, and suffering from major depressive disorder or attention deficit hyperactivity disorder (ADHD). In addition, the strong claim is an important component of the arguments that abortion and incarceration reduce crime. We will examine the abortion controversy in this section, but we defer looking at incarceration to [Section 23.5](#).

A much larger number of studies examine what we call the *weak claim of criminogenic characteristics*, that relatively stable traits predict who commits crimes. These studies show only that people with certain characteristics are more likely to commit crimes than people without these characteristics, and sometimes that the relationship between having these traits and committing crimes is causal. We know only, for instance, that graduates of some high-quality preschools commit fewer crimes; we do not know if the total number of crimes goes down (with an appropriate lag) when such preschools expand. Someone else could commit the crimes that these preschool graduates would have committed.

The weak claim is about individual behavior: who commits crimes? The strong claim is about general equilibrium: how many crimes are committed? To take a labor market analogy, a weak claim would state that high school dropouts are more likely to work at fast food restaurants than otherwise similar workers who are better educated, while a strong claim would state that an increase in the proportion of dropouts in the population raises the proportion of workers who are employed at fast food restaurants.

The final part of this section looks at disparities in crime across identity groups and asks how much of the concentration of crime in particular communities can be explained by their distinctive distribution of characteristics.

One way to distinguish this section from the next is to say that this one is about preferences and endowments, while the next is about incentives. That is approximately true. But, in fact, neither the strong claim nor the weak claim articles establish that the traits in question are operating through preferences rather than incentives. For instance, people with a history of ADHD are more likely to commit violent crimes; there are both strong and weak claim articles on this relationship. But people with a history of ADHD might be more likely to commit these crimes not because they are more aggressive or impulsive, but because they earn less money and so would suffer a smaller loss if they were imprisoned. Almost any trait that changes preferences in the direction of more criminal activity (or more current consumption) is likely to be correlated with low income, if only because most employers do not like their workers committing crimes and being arrested. Characteristics might also alter the nonpecuniary costs of imprisonment: a person unable to enjoy a happy and fulfilling life outside prison might lose less from entering prison. We do not think that any of the articles we survey definitively separate preferences from

incentives, although [Lochner and Moretti \(2004\)](#), discussed further below, shed considerable light on this question for education.

23.2.1 The strong claim of criminogenic characteristics

23.2.1.1 Age and gender

About 93% of state and federal prisoners in 2010 were male, as were 87% of residents of juvenile detention facilities in 2011, and 87% of jail inmates in 2011. In 2011, 64.2% of arrestees were between 15 and 34 years old, compared with 27.4% of the residential population.⁵ Hence, the weak claim about young men is undoubtedly true. Evidence for the strong claim, however, is not consistent, and many studies fail to support it. [Pratt and Cullen \(2005\)](#) provide a meta-analysis of over 200 ecological studies of crime, most of which contain age structure and sex ratio variables. Of 31 predictors that they study, age structure and sex ratio are ranked 16th and 19th most powerful, respectively. [Pratt and Cullen \(2005, p. 399\)](#) describe them as “mid-range predictors of crime [that] are likely to make a significant contribution to the proportion of explained variance in a statistical model” but certainly not as robust as the effects of predictors ranked higher in the distribution. Perhaps the results would have been stronger had more studies looked at the proportion of young men in the population, rather than age and sex separately.⁶

23.2.1.2 Childhood lead exposure

The evidence that atmospheric lead affects the volume of crime is more convincing than the evidence regarding age and gender. Childhood lead exposure increases impulsivity and aggressiveness, and lowers IQ. Animal studies show links to ADHD and antisocial activity. Following the 1970 Clean Air Act, lead was almost entirely removed from gasoline in the United States between 1975 and 1985. Among children under 6 years, lead in blood fell from 18 $\mu\text{g}/\text{dL}$ in 1976 to 2.8 $\mu\text{g}/\text{dL}$ in 1991. Lead was phased out of gasoline in different states at different times, and [Reyes \(2007\)](#) uses this interstate variation in lead reductions to identify the effect of childhood lead exposure on crime. She uses state-year observations to link crime to childhood lead exposure with a lag of 20–30 years. The elasticity of violent crime (essentially, robberies plus aggravated assaults) with respect to lead exposure is around 0.8. Lead has no significant effect on property crime. The effect on murder is not robust, but is suggestive. Since gasoline use per square kilometer and the number of children per square kilometer are both higher in urban areas than other areas, the effects of leaded gasoline are probably concentrated in urban areas.

⁵ Sourcebook of Criminal Justice Statistics, tables 6.33.2010, 6.11.2011, 6.17.2011, and 4.4.2011.

⁶ Two studies of how crime reacts to unconventional closings of high schools, [Jacob and Lefgren \(2003\)](#) and [Luallen \(2006\)](#), come close to supporting the strong claim by showing that juvenile property crime increases when high schools are closed; they would have supported the strong claim if their dependent variable had been total property crime, rather than juvenile property crime.

This finding is like the other supporting results about the strong claim. Instead of moving directly from changes in the distribution of characteristics in the population to the volume of crime, these studies start with some exogenous event (in this case, reduction in atmospheric lead) that plausibly changes the distribution of characteristics in the population (fewer people with ADHD), and show that the exogenous event causes changes in the volume of crime.

23.2.1.3 Schizophrenia, ADHD, and major depressive disorder

“Mental illness” refers to many different conditions, and “crime” refers to many different activities. Some mental illnesses (e.g., substance abuse disorder and conduct disorder) are even crimes by definition. It makes more sense to try to link particular disorders to particular crimes than to try to link two amorphous abstractions.

We are aware of only two articles that provide evidence for a strong claim. [Marcotte and Markowitz \(2011\)](#) look at changes in prescriptions for antipsychotic drugs for treatment of schizophrenia, various classes of antidepressants for treatment of major depressive disorder, and stimulants for treatment of ADHD. These psychopharmaceuticals are all known to be effective. Marcotte and Markowitz use state-quarter variation in prescriptions to identify the effects on violent crime, with state and quarter fixed effects and many covariates. Prescriptions for stimulants and for “newer-generation antidepressants”—bupropion is the best known drug in this class—reduced violent crime significantly. The elasticities were small, less than 0.1, but were economically significant because of the fast growth in the number of prescriptions. Antipsychotics and the other tested class of antidepressants—serotonin reuptake inhibitors and serotonin-norepinephrine reuptake inhibitors—did not significantly reduce violent crime. Along similar lines, [Cuellar and Markowitz \(2007\)](#) show that more state Medicaid spending and more Medicaid prescriptions for stimulants are associated with less violent crime; and so is more state spending on older antidepressants.

Evidence for the weak claim, that people with various kinds of mental disorders are more likely to commit various crimes, is abundant; see, for instance, [Frank and McGuire \(2011\)](#). People with mental illness, moreover, are more likely to be victims of crime. A review of the older literature on this question concluded: “Victimization is a greater public health concern than perpetration” ([Choe et al., 2008](#), p. 153). Thus it is unclear how much the mechanism that leads from more psychopharmaceutical prescriptions to fewer violent crimes acts by changing the distribution of characteristics of potential offenders, and how much it operates by changing the distribution of characteristics of potential victims.

23.2.1.4 Family structure

People who grow up in families that do not always have two parents are more likely to commit crimes, and [Antecol and Bedard \(2007\)](#) show that this relationship is probably causal (they use state changes in divorce laws as an instrument). Many cross-section

studies also find that the proportion of families headed by single parents is associated with more crime; in fact, [Pratt and Cullen \(2005\)](#) find that family disruption is the ninth strongest of the predictors of crime in their meta-analysis. Cross-sectional studies often regress crime on contemporaneous measures of family structure, which is not the relevant variable. However, [Kendall and Tamura \(2010\)](#) find that more nonmarital births in a state are associated with increases in the rates of murder and car theft after 15–20 years. To our knowledge, no test of the strong claim has been attempted with a strategy for identifying causality.

23.2.1.5 Abortion

[Donohue and Levitt \(2001, 2004, 2008\)](#) famously—and controversially—argue that legalizing abortion in the United States in the early 1970s caused a large portion of the crime decline in the 1990s. We examine this claim here because the mechanism they use to explain this association relies on the strong claim of characteristics. The argument is based on two propositions: first, that legalizing abortion caused a major change in the distribution of characteristics of young men, and second, that this change in the distribution of characteristics reduced crime. Only the second proposition is relevant for this section, since it is an instance of the strong claim of characteristics. But unless the first proposition is correct (abortion changed characteristics), abortion tells us nothing about the second proposition (characteristics changed crime). We do not believe that the first proposition is supported by the available evidence.

The naïve theory about abortion and characteristics is that the number of births is reduced one-for-one by the number of abortions: nonaborted births are unaffected, and the change to the number of children with any characteristic is just the number of aborted children who would have had that characteristic had they been born. Since women who have abortions in the United States plausibly would have had children with more criminogenic characteristics, including unwantedness, than pregnant women who do not, this theory implies that abortion produces a less criminogenic distribution of characteristics—after a lag of about 15–20 years. Donohue and Levitt implicitly refer to this theory as motivation for their work, and it underlies the specifications they estimate.

The naïve theory does not describe the US experience with abortion. The consensus estimate is that legalizing abortion reduced the number of births by around 4% or 5% ([Levine et al., 1999](#); [Ananat et al., 2009](#)). But the number of abortions, relative to the number of births, is an order of magnitude greater: around 30% or 40% in the 1970s ([Statistical Abstract, 2001, table 92](#)). Legalized abortion increased the number of pregnancies—[Ananat et al. \(2009\)](#) show causality—and decreased the ratio of the number of births to the number of pregnancies. One can imagine not only babies who were never born because of abortion, but also babies who were born *because* of abortion, and who would not exist had abortion not been legal.

The available evidence indicates that abortion did not substantially alter the distribution of any characteristics within cohorts, although it made cohorts smaller to a modest degree. Ananat et al. (2006, 2009) try to find effects on a number of noncriminal outcomes; while they find modest effects for a few of these, Joyce (2009) seems accurate in his assessment that they do not show large or pervasive changes in noncriminal outcomes for cohorts affected by abortion law changes. On criminal outcomes, both Ananat et al. (2006) and Joyce (2009) show that abortion did not reduce arrest rates, although total arrests declined because cohorts were smaller.

The clearest change associated with the legalization of abortion is a decline in the size of cohorts by about 4–5%. As we saw in Section 23.2.1.1, the effect of age structure on aggregate crime has been studied extensively, and does not seem to be very important. If abortion were a clean natural experiment that exogenously reduced cohort size by a large amount, it could tell us something about this important effect. But the saga of the 1970s is anything but a clean natural experiment; *Roe v. Wade* was not the Mariel boatlift in reverse.

Donohue and Levitt (2001, 2004) do show that drops in crime by state follow increases in abortion rates with appropriate lags, but do not establish that this relationship is causal. We do not understand why aggregate crime correlates with effective appropriately lagged abortion rates, since the explanation offered by Donohue and Levitt, and the one that most immediately comes to mind, is not tenable. The key may be that abortion rates are not exogenous, as Ananat et al. demonstrate, and the forces that were moving abortion rates may have had other effects as well (on the behavior of potential victims as well as that of potential offenders). Feminism seems a likely candidate explanation: it obviously moved abortion rates in the 1970s, and it has had innumerable, pervasive, and long-lasting repercussions. How that might have worked and how it could be tested, we have no idea. We hope future scholars will approach the Donohue and Levitt correlation as a puzzle, not a battleground.

23.2.2 The weak claim of criminogenic characteristics

Many characteristics appear to predispose people to commit crimes. In this section, we look only at areas that have received considerable attention: genetics, education, family structure, social and behavioral skills, military service, and peers. (We have already noted that the weak claim holds for age, gender, and some psychiatric disorders.)

In each of these areas, certain traits are associated with later evidence of criminal activity, usually self-reports, arrests, or convictions. So weak claim articles show more direct evidence of criminal *offending* than strong claim articles do—in strong claim articles, after all, greater crime could be caused by changes in the behavior of potential victims or the police. But even the weak claim articles do not show actual

offending: the traits might alter the probability of self-reporting, being arrested, or being convicted, conditional on offending.⁷

23.2.2.1 Education

Almost all work on education has been targeted on the weak claim. The weak claim has been demonstrated for some kinds of preschool education, for school quality, and for educational attainment, especially high school graduation. These processes seem to operate not through the cognitive skills of the students who experience them, but through other, noncognitive, channels.

Several randomized controlled trials (RCTs) have been completed on high-quality preschool programs, and some of them (but not all) have shown that participants in these programs are less likely to be punished for crimes when they are older. The High Scope Perry Preschool in Michigan provided half-day preschool for 3- and 4-year-olds in the early 1960s, along with biweekly home visits. By the age of 40 years, it had reduced the fraction of men arrested five or more times from 0.69 to 0.45, and the fraction of women from 0.34 to 0.24. It also raised incomes at age 40 years. The Syracuse University Family Development Program reduced juvenile delinquency by age 15 years from 22% to 6%. On the other hand, two other RCTs involving similar programs—the Abecedarian Project in North Carolina and the Infant Health and Development Program at multiple sites—failed to reduce arrests significantly, although participants have been observed only to ages 21 and 18 years, respectively (Lochner, 2011).

Quasi-experimental studies, some on much larger programs, often show reductions in crime too, but there are some exceptions. The Chicago Child Parent Center was evaluated by comparison with a matched sample from nearby neighborhoods that would have been geographically ineligible for the program. It reduced arrests by one-third by age 18 years (Reynolds et al., 2001). Three other studies—Garces et al. (2002), Carneiro and Ginja (2008), and Deming (2009)—look at the large federally funded preschool program called Head Start. The first two articles find significant reductions in being booked, being charged with a crime, or being sentenced, especially for African Americans, while the third article does not find a significant impact on crime.

There is no explanation known now for why these programs had disparate results. None of these programs raised teenage measures of cognitive skill, and the results on high

⁷ If certain traits are correlated with committing crimes, Bayesian juries will be more likely to convict people with those traits, conditional on the evidence presented in court, and Bayesian police will be more likely to arrest them. Juries and police will also act this way if they believe the correlation is present, but are wrong. Thus, in the experiment by Valla et al. (2011) that showed college students being able to identify convicted criminals from head shots, the mechanism may have been that juries convicted people whom they thought looked like criminals, and college students had the same beliefs as jurors. That is, students may have identified those more likely to be convicted and not necessarily those more likely to have been offenders.

school graduation rates are mixed but only weakly correlated with the results on crime. Interventions with young children can reduce crime in adolescence and adulthood, but we do not know yet precisely what those interventions are, or through which mechanisms they operate.

School quality also matters: children who attended better schools are less likely to commit crimes in adolescence and young adulthood. Cullen et al. (2006) and Deming (2011) both examine the results of public school choice lotteries, in Chicago and Charlotte-Mecklenburg, respectively. Winning a lottery increases the probability that a student will attend a public school that is better along several dimensions. In neither case did winning the lottery improve academic outcomes, but both articles find that lottery winners were less likely to be arrested or incarcerated—during high school for Cullen et al., and after high school for Deming.

Next consider educational attainment. Less educated people are more likely to commit crimes than more educated people. In 1997, two-thirds of prison inmates in the United States were high school dropouts (Lochner, 2011). The association is not necessarily causal: some other factor—time preference or poor families, for instance—could be causing people to be both dropouts and criminals. By using the legal school leaving age as an instrument, researchers have fairly well quite convincingly established that the relationship is causal: low educational attainment (in particular, not finishing high school) causes criminal activity. Lochner and Moretti (2004) were the first to establish this relationship for the United States, and Machin et al. (2011) obtained similar results for the United Kingdom.⁸

This effect probably operates through earnings—thus incentives—rather than through preferences. Lochner and Moretti (2004) use the existing literature to compound the effect of education on earnings with the effect of earnings on crime, and thus compute the effect of education on crime that would be expected if education did not affect preferences. This estimate is not very different from the effect that they find. Similarly, when Grogger (1998) enters both education and earnings in a cross-sectional regression to explain individual criminal activity, the coefficient on education becomes insignificant.

While evidence for the weak claim for education is strong, evidence for the strong claim is weak. In a cross-national study of murder and robbery, for instance, Fajnzylber et al. (2002) find no significant effect for average educational attainment. Of the 31 predictors of crime in the meta-analysis by Pratt and Cullen (2005), education is the 29th strongest, with essentially no effect. An exception is the study of Buonanno and Leonida (2006), who use generalized method of moments methods to look at property crime in a panel of Italian regions.

⁸ These articles are about the weak claim rather than the strong claim because they show only that crime committed by cohorts affected by the legal change went down, not that total crime went down.

23.2.2.2 *Psychological and character traits*

People who approach life in certain ways are more likely to commit crimes than people who do not. Anger, hostility, and impulsiveness are positively correlated with conduct problems; agreeableness (straightforwardness, compliance, altruism, and trust) and conscientiousness are negatively correlated. Personality matters.

More importantly, the relationship between personality traits and crime appears to be causal. We know this from a series of RCTs that have shown that interventions designed to alter how people approach life can reduce crime.⁹ But not every personality-altering intervention reduces crime. The most famous interventions are all failures, including Scared Straight, Drug Abuse Resistance Education (DARE), and boot camps (Hill et al., 2011).

Which traits are altered by crime-reducing interventions is not settled. Moffitt and Ross (2011) argue that self-control is the key trait. Hill et al. (2011) take a more eclectic approach, but emphasize the role of developing “more effective strategies for dealing with social situations.” Interventions that get young people to change their routine or automatic reactions to possibly stressful social situations often reduce crime (but whether they make young people who change better off is unknown).

Becoming a Man (BAM) was an RCT in Chicago that showed impressive crime reductions from trying to change the social skills of teenage boys. The premise was that many teenage boys in poor neighborhoods reacted to stressful social situations in ways that led to crime: they were hypervigilant to threat cues, and quick to attribute malevolent intent to others (they exhibited “hostile attribution bias”). BAM combined exposure to prosocial adults with cognitive behavioral therapy in order to inculcate less criminogenic habits. The intervention was short and inexpensive. Treatment reduced arrests for violent and other crimes (mainly vandalism, trespassing, and weapons offenses) by 30–40% in the first year. Academic outcomes also improved (Heller et al., 2013).

After the first year, crime reductions were not statistically significant. Maybe this “fade-out” should not be surprising. If the neighborhoods in which these boys lived taught them to respond with hostile attribution bias before BAM, these neighborhoods might be expected to teach the same thing after BAM. Previous interventions that tried to inculcate social skills were not as successful as BAM (Hill et al., 2011), and Heller et al. (2013) speculate that BAM’s key ingredient was cognitive behavioral therapy. But the key ingredient may have been BAM’s sample size and its pervasiveness. There were 2740 participants, about three-fourths of the male students in the schools, in the program. For a while anyway, the BAM treatment group may have encountered other members of the BAM treatment group often enough that the new habits would not be unlearned

⁹ Of course, these traits do not cause “crime” in general; they cause those crimes that people in the RCTs are most likely to get caught committing or admit to. Impulsiveness, for instance, may not make people better at embezzling, although it may make them more likely to get caught if they are embezzling.

quickly. Moreover, the observed “fade-out” may not have been fade-out at all, but may have resulted from partial learning of new habits by the control group, as encounters between treatment and control group members led both to a new equilibrium.

23.2.2.3 Brain functioning, brain structure, in utero experience, and genetics

On average, the brains of people who commit crimes, especially violent crimes, tend to work differently from those of people who do not. The brain structures also tend to be different, and this is part of the reason for different brain functioning. Different brain structure, in turn, is associated with postbirth experiences (e.g., exposure to poor parenting and atmospheric lead), with *in utero* experience, and with genetics. A few of these relationships have been shown to be causal. Lombroso did not miss the mark entirely.

Just as crime covers a large array of activities and motives, the ways that criminals differ from noncriminals are many and varied (Raine, 2013). One set of differences is tied to the cognitive parts of the brain such as the ventral prefrontal cortex and the angular gyrus: those who commit certain types of crimes fail to regulate their emotions well and often make bad judgments. But crime is not solely the failure of rationality to control passion. The limbic, emotional parts of the brain and the autonomic nervous system also matter: criminals tend to be fearless and sensation seeking (this is tied to a slow heartbeat); they also lack empathy and moral reasoning skills (fear conditioning is tied to developing a conscience). Parts of the cortex that handle inhibition and the ability to learn to avoid punishment also tend to differ in criminals.

In all of these areas, differences in both functioning and structure have been detected. For instance, in one experiment, 41 murderers and 41 age- and sex-matched normal controls had their brains scanned after they had performed a repetitive task that required concentration. The murderers, especially the reactive emotional ones, showed significantly less prefrontal glucose metabolism; their brain *function* was different (Raine, 2013, pp. 66–67). The brain function of pedophiles has also been shown to differ from that of people sexually attracted to adults (Ponseti et al., 2014). *Structure* studies have shown that in people in whom antisocial personality disorder has been diagnosed there is around 11% less gray matter (by volume) than average in their prefrontal cortexes, and that psychopaths have 18% less amygdala volume. The parts of the amygdala with the greatest volume reduction govern fear conditioning and avoidance learning (Raine, 2013, pp. 139, 161–162). Almost all of the studies of brain functioning and structure say nothing about causality, but some of them measure physiology before crimes are committed: for instance, in Mauritius, lack of fear conditioning at age 3 years predicted criminal convictions by age 23 years (Raine, 2013, pp. 118–119).

What causes these differences in function and structure? Obviously, environment matters. Most of the experiences we have already shown to influence criminality—breathing atmospheric lead, using psychotherapeutic drugs, going to school, being raised by a lone parent—probably operate through changes in brain function and structure. Indeed, the best evidence for a causal connection between brain structure and behavior

comes from bizarre and gruesome “natural experiments” that destroy parts of a person’s brain and alter his or her behavior. The most famous neuroscience natural experiment was performed on Phineas Gage. He was a shrewd, smart, reliable, self-controlled railway construction foreman until one afternoon in 1848, when a blasting accident sent a metal tamping rod, three and half feet long and more than an inch in diameter, through his skull, entering his lower left cheek and exiting from the top middle part of his head. Although Gage survived, the accident probably destroyed much of his prefrontal cortex, and he became profane, impatient, impulsive, irresponsible, sexually promiscuous, and unable to hold a steady job (Raine, 2013, pp. 143–145).

In utero experience also affects brain function and structure. In particular, maternal smoking and alcohol consumption are correlated with abnormal brain development in the fetus and later criminal offending. The smoking relationship holds for second-hand smoke, and the alcohol relationship holds for children adopted away from their natural mothers (Raine, 2013, pp. 198–205). Hunger during pregnancy also causes children to develop antisocial personality disorder; this is a result from a natural experiment from the Dutch “hunger winter” of 1944–1945, as reported by Neugebauer et al. (1999).

Finally, genetics also matters. Twin studies and adoption studies both show correlations between parent offending and child offending, but of course no causality. So far several genes have been found to influence brain structure, brain functioning, and aggressive behavior and delinquency. Of these, *MAOA* is the most controversial and best studied. The *MAOA* gene produces the enzyme called monoamine oxidase A (MAOA). This enzyme enters into the operation of serotonin, and several other neurotransmitters. When nerve cells transmit information between each other, they produce serotonin in the gaps between them. If the serotonin remains outside the cells, it becomes toxic. MAOA works to degrade the serotonin, and get it reabsorbed into the cells. Poor serotonin reuptake may lead to ADHD, drug abuse, alcoholism, and impulsivity. MAOA may also contribute to reduced volume of the amygdala and the orbitofrontal cortex (Raine, 2013, pp. 61, 260).

There are several different genetic polymorphisms of the *MAOA* gene—that is, people inherit different versions of the instructions for producing MAOA. A few people (one famous Dutch family in particular) produce no MAOA. A tiny number of people have the two-repeat version. Most people have three-repeat, 3.5-repeat, four-repeat, or five-repeat versions. The more repeats, the more enzyme produced. Most studies divide the population into a “low-*MAOA*” group—that is, those with two-repeat and three-repeat *MAOA* versions—and a “high-*MAOA*” group—everybody else.

The main consensus of many studies (though not a unanimous one) is that among English-speaking men of predominantly European descent, low-*MAOA* individuals who suffer abuse of one kind or another in childhood are more likely to engage in anti-social activities and violent crime in young adulthood than high-*MAOA* individuals who experience abuse. The intuition behind this result is that childhood abuse leads to heavy serotonin release, which damages nerve cells unless copious amounts of MAOA are on

hand. This result, however, does not appear to hold for Taiwanese (Lu et al., 2003), or African Americans (Widom and Brzustowicz, 2006).

The proportion of people with the low-*MAOA* genotype ranges from 30% to 80% in various populations. In that sense, it is a weaker predictor of violent behavior than being male. Furthermore, one should also be careful in interpreting correlations between genotypes and behavior. People with a certain genotype are not randomly assigned to parents. The measured effect is that of having the genotype, and of having parents typical of the parents of people who have the genotype.

23.2.2.4 Other characteristics

Galliani et al. (2011) find that men who have served in the military are more likely to commit crimes. They use the results of the draft lottery in Argentina to identify a causal effect of military service. This effect is present in both war and peace.

People who eat more fish are less violent and hostile. This has been demonstrated in many studies, including RCTs. The active ingredient in this relationship appears to be omega-3 fish oils, which enhance the working of brain cells in various ways. In a cross section of developed nations, there is a negative correlation between seafood consumption per capita and number of homicides per capita. One might therefore claim that fish eating satisfies the strong claim of characteristics as well as the weak claim, but the evidence is not abundant (Raine, 2013, pp. 213–216).

23.2.3 Race and ethnicity

We have omitted race and ethnicity from our discussion so far because they are correlated with far more than preferences (if they are correlated with preferences at all). But they are characteristics that are correlated with crime.

As far as the strong claim goes, the meta-analysis of ecological studies by Pratt and Cullen (2005) found “percent nonwhite” to be the fourth best predictor of aggregate crime out of 31 variables tested; it has a far stronger correlation with aggregate crime than any of the variables we have discussed. Since most of these studies include many other variables, we can be fairly certain that the effect of race on aggregate crime is not explained by the traditional variables that we found could affect aggregate crime—age structure and single parenthood.

What about the nontraditional characteristics for which the strong claim holds—lead exposure, ADHD, major depressive disorder, *in utero* experience? Minorities may plausibly be exposed to more atmospheric lead than non-Hispanic whites, but the relationship has not been explicitly explored. Reyes (2007), however, argues that lead exposure affects crime primarily through ADHD (the IQ channel is minor).

Hence we are left with ADHD and major depressive disorder. These disorders are unlikely to explain much of the impact of race on aggregate crime because the prevalence is fairly similar across groups; indeed, these disorders appear to affect non-Hispanic whites

disproportionately. Diagnosed ADHD is slightly more prevalent among non-Hispanic white children (age 5–17 years) than among non-Hispanic black or Hispanic children (Akinbami et al., 2011). Kessler et al. (2006), in a large survey of clinician-assessed adult ADHD, found that non-Hispanic whites were more likely to have this disorder than minorities; this relationship remained after a number of other variables had been held constant. Riolo et al. (2005) report greater prevalence of clinician-assessed major depressive disorder among whites than among African Americans or Mexican Americans.

The weak claim also holds for race and ethnicity: minorities are more likely to be arrested and incarcerated for most crimes (and possibly to commit more crimes). Education does not explain the disparity: Lochner and Moretti (2004) estimate that educational attainment accounts for only about 23% of the difference in incarceration rates between blacks and whites. Minorities, moreover, are not more likely to be veterans than non-Hispanic whites are. In 2010, 11.3% of veterans were African Americans, as opposed to 13.1% of the general population and 12.2% of the population 18 years and older. Similarly, 5.5% of veterans were Hispanic, as opposed to 16.9% of the general population and 13.6% of the population 18 years and older.

The argument that characteristics related to preferences might be responsible for a sizable portion of the racial crime gap is probably strongest for the personality traits such as aggressiveness, impulsiveness, and hostile attribution bias that we discussed in Section 23.2.2.2. These are not clinical disorders, and no good national prevalence data seem to be available. They were common, however, among the young black men in Chicago for whom BAM was designed.

What is not clear about these characteristics is whether they should be treated as fundamentals that are acquired genetically or haphazardly early in life, or as best responses to the environment in which many black youths find themselves. The quick fade-out of the effects of BAM suggests the latter.

Genetic explanations have also been examined, although little genetic research has involved minority groups. The prevalence of low-*MAOA* genotype varies by race: it is present in about 30% of white males, 56% of Maori males, and 77% of Chinese males (Raine, 2013), and it appears to be present in African Americans more than in whites in some samples—for example, Sabol et al. (1998) found that 59% of African Americans and 33% of whites had the low-*MAOA* genotype, although Widom and Brzustowicz (2006) found that in both white and nonwhite males the low-*MAOA* genotype was present about 41% of the time; both samples had about 100 nonwhites or African Americans, but Sabol et al. had a much larger sample of whites. It is hard to see how *MAOA* genotype differences could explain a large proportion of the black-white crime gap, since *MAOA* appears to have little effect on violence among African Americans, since ADHD prevalence does not vary by race, and since persons of Chinese descent in the United States tend to be less violent than whites. Genetic variation may be important, but genetic variation depends not just on *MAOA*, but on many other genes as well.

To summarize, the characteristics we have discussed in this section seem not to be very important causes of the racial crime gap, and it is important to consider incentives and interactions as determinants of behavior. We do this in the sections that follow.

It is apparent that criminogenic characteristics are neither evenly distributed across space nor stable over time. Atmospheric lead is a clear example of extreme spatiotemporal variation, being concentrated in urban areas and declining dramatically over time. The sorting of households across residential locations can also induce variations in population characteristics across space. But such sorting is at least in part a response to incentives, which we examine next. We revisit characteristics in [Sections 23.6](#) and [23.7](#), when looking at variations across time and space in the pattern of crime.

23.3. INCENTIVES AND DETERRENCE

Instead of looking at who criminals are in order to understand why they commit crimes, another tradition in criminology looks at the incentives they face. This tradition starts with Cesare Beccaria's (1738–1794) treatise *Dei Delitti e Della Pene* (“On Crimes and Punishments”). Beccaria believed that people rationally commit crimes when it is in their best interest to do so, and that they could and should be deterred by expected punishments of appropriate size. He was the first economist to study crime.

Beccaria's view has both positive and normative components. The positive implication is that differences in rewards and punishments can explain differences in criminal offending over time and space. The normative implication is that governments should deter crime by setting punishments for various crimes as high as the gains that criminals might expect from committing those crimes, and no higher. For Beccaria, the relevant question in deciding whether a person should be punished is whether that person has committed a crime, not whether that person will commit a crime. The criminal justice system is backward-looking; prediction, the central concern of the previous section, plays no role.

Two centuries later, [Becker \(1968\)](#) revived Beccaria's approach: he wrote that his “efforts can be viewed as a resurrection, modernization, and thereby I hope improvement on these much earlier pioneering studies” ([Becker, 1968](#), p. 209). Becker's article inspired a continuing stream of research among economists, and his name has become synonymous with deterrence, in particular, and incentives to commit crime, in general.

Although dividing approaches to crime into those that emphasize characteristics ([Section 23.2](#)) and those that emphasize incentives (this section) is convenient for expositional purposes, the distinction is not a deep unbridgeable gap, but a question of emphasis, and emphasis depends on the location of heterogeneity. If, for instance, we are looking at a cross section and people differ little in criminogenic characteristics but face different incentives, perhaps because of more effective policing or harsher punishments in some localities, then incentives will predict crime well and criminogenic characteristics will not. Conversely, if everyone faces the same expected punishment for committing a crime, criminogenic characteristics will determine who commits crimes and who does

not. The same holds for the time series: whether changes in crime over time can be explained better by changes in incentives or changes in criminogenic characteristics depends in large measure on which is changing more during the period in question.

Normative prescriptions also depend on the location of heterogeneity. [Beccaria \(1764, p. 101\)](#), for instance, argues that the appropriate expected punishment for a crime is precisely the gain that criminals realize from committing it: anything less is ineffective, and anything more is unjust and welfare destroying. This formulation, however, presumes no heterogeneity in the gains that people might realize from committing identical crimes.¹⁰ Similarly, universal preschool may be very effective per dollar or per student through reducing criminogenic characteristics if students are homogeneous and the same “dose” reduces criminality in everyone, but if students are heterogeneous and many are in no danger of committing crimes with or without preschool, the attractiveness of this strategy is considerably diminished.

In this section, we review theoretical and empirical results on deterrence, and on incentive effects of earnings, unemployment, and poverty.

23.3.1 Certainty and severity

The effectiveness of a threat of punishment depends on how certain it is that punishment will follow the crime, and how severe that punishment is. It has been traditional to think of certainty and severity as two separable dimensions of punishment, and to try to estimate the effects separately. In fact, the two dimensions are usually not entirely separable, either in policy or in estimation, but the distinction is still useful. [Beccaria \(1764, chapter 19\)](#) added a third dimension, celerity, arguing that swift punishments are a greater deterrent than delayed ones, but modern economics has not followed him.

Empirical studies of severity appear to be harder than studies of certainty. The primary punishment for serious crimes in developed countries is incarceration, and incarceration can reduce crime through incapacitation as well as deterrence (prisoners cannot commit index crimes against nonprisoners, and interprisoner crime is hard to commit). Execution, a less widespread mode of punishment, also incapacitates (more effectively than incarceration). A host of ancillary punishments, meted out by both governments and private parties, are also used and are little studied.¹¹ Some, such as disenfranchisement, have little incapacitative effect, while others, such as restrictions on the residences and activities of former sex offenders, are intended to be primarily incapacitative. Ancillary punishments such as ineligibility for student loans and occupational bars do not have incapacitative effects; indeed, they may increase the incentives for crime after imprisonment.

¹⁰ It is possible, of course, to vary penalties on the basis of offender characteristics such as income or wealth, and many countries in Europe have implemented such contingent fines for traffic violations.

¹¹ See [Curtis et al. \(2013\)](#) for a survey of how public housing authorities punish people who have been arrested for or convicted of various crimes, and their families.

It is commonly believed that crime responds more to certainty than to severity—specifically, that the elasticity of crime with respect to measures of certainty is greater (in absolute value) than the elasticity with respect to severity. For instance, [Beccaria \(1764, p. 100\)](#) wrote: “Crimes are more effectually prevented by the *certainty* than by the *severity* of punishment. . . The certainty of a small punishment will make a stronger impression, than the fear of one more severe, if attended with the hopes of escaping; for it is the nature of mankind to be terrified at the approach of the smallest inevitable evil, whilst hope, the best gift of Heaven, hath the power of dispelling the apprehension of a greater; especially if supported by examples of impunity, which weakness or avarice too frequently afford.”

[Becker \(1968\)](#) gives two separate circumstances that make the certainty elasticity greater than the severity elasticity, one concerning choice by individual potential offenders, and the other concerning strategies of policymakers.

For individual potential offenders, the certainty elasticity will be greater than the severity elasticity when they are “risk preferrers” in wealth equivalents. For Becker, penalties are measured in wealth equivalents, as if they were all fines. Thus, a person’s expected utility EU_c after committing a crime is the expected value of a lottery:

$$EU_c = pu(\gamma - f) + (1 - p)u(\gamma),$$

where p is the probability of being convicted and punished, $u(\cdot)$ is a von Neumann–Morgenstern utility function, γ is wealth if not convicted and punished, and hence f is the wealth equivalent of whatever penalty is meted out. Becker interprets p as certainty and f as severity.

Consider an increase in p adjusting f to hold pf constant; that is, consider equal percentage changes in certainty and severity, but in opposite directions. Write $pf = k$. Then

$$EU_c = pu\left(\gamma - \frac{k}{p}\right) + (1 - p)u(\gamma),$$

and so

$$\frac{\partial EU_c}{\partial p} = -[u(\gamma) - u(\gamma - f)] + fu'(\gamma - f).$$

By the mean value theorem, there is some $z \in [\gamma - f, \gamma]$ such that

$$fu'(z) = u(\gamma) - u(\gamma - f).$$

Hence,

$$\frac{\partial EU_c}{\partial p} = f[u'(\gamma - f) - u'(z)].$$

For a person who is risk preferring, $u'' > 0$. So $u'(z) > u'(\gamma - f)$, since $z > \gamma - f$. In this case, increasing p with an offsetting decrease in f reduces the expected utility of crime.

The opposite is the case for people who are risk averse. Crime will respond more to changes in severity than to equal percentage changes in certainty.

For policymakers, Becker shows that if p and f are chosen optimally, then in the neighborhood of the optimum, the elasticity of crime with respect to p is greater than the elasticity with respect to f . The argument is a *reductio ad absurdum* in the spirit of the proof that a profit-maximizing monopolist never operates on the inelastic portion of the demand curve.

For Becker's policymaker, the optimum minimizes the sum of three costs: the costs of crime, the costs of detection, and the cost of punishment. For any fixed amount of crime, the optimal pair (p, f) minimizes the sum of the cost of detection and the cost of punishment. Given the amount of crime, detection costs depend on p —the more police, the more crimes solved—and punishment costs depend on the product pf . Suppose (p^*, f^*) minimizes this sum, and the severity elasticity is greater than or equal to the certainty elasticity: at the margin, potential criminals are risk neutral or risk averse. Then reduce p and increase f to keep pf constant. This holds punishment costs constant and (weakly) reduces crime (crime is constant if the marginal potential criminals are risk neutral). But then detection costs can be lowered without raising the initial level of crime. Hence, (p^*, f^*) cannot be optimal.

Essentially, with risk-averse criminals, policymakers can drive crime and enforcement cost to zero by threatening an infinitely horrible punishment at infinitesimally low probability. Hence, if (p, f) is set optimally, the potential criminals on the margin are risk preferring, and certainty elasticity exceeds the severity elasticity.¹²

There is little reason to believe that either of Becker's two conditions, risk-preferring marginal criminals or optimal policies, is satisfied at any historical moment. But inequality of the two elasticities should not be taken as a sign of suboptimality. Only by assessing evaluating costs and elasticities in a particular situation can economists determine assess the optimality of particular constellations of policies.

23.3.2 Empirical studies of deterrence

Since Becker's work, many estimates of deterrence have been made. [Durlauf and Nagin \(2011\)](#) provide a complete and critical review. We will focus on the most compelling work. The bottom line is that substantial evidence finds large effects of certainty of punishment on crime, but evidence regarding severity is spotty.

¹² Note that this argument requires potential criminals to make accurate assessments of the detection probability. In contrast, Beccaria's understanding of the trade-off between certainty and severity hinges on an underestimation of the likelihood of punishment relative to its true value, based on exaggerated "hopes of escape." If this effect is sufficiently strong, then even those who exhibit risk aversion conditional on their subjective beliefs may be better incentivized to abstain from crime through certain rather than severe punishments. Put differently, the certainty elasticity may be higher than the severity elasticity even if the marginal criminal is risk averse, provided that the probability of apprehension is systematically underestimated.

23.3.2.1 *Certainty studies*

Whether a crime is punished depends mainly on how police respond to it, and so studies of certainty are almost all studies of police and what they do. Courts matter too (and for courts, both type 1 and type 2 errors affect deterrence), but we are not aware of any studies of court error.

Responses to terrorist attacks

Various terrorist incidents and threats have caused police departments to redeploy their forces in unexpected ways. These incidents are plausibly exogenous to the more mundane index crimes that we study, and the redeployments are natural experiments for the effect of police presence on index crime.¹³

Di Tella and Schargrodsky (2004) look at a redeployment in Buenos Aires that followed the bombing of a Jewish center in 1994. Police were stationed at every Jewish and Muslim institution. Car thefts fell by 75% on the blocks where the extra police were stationed, and did not rise elsewhere. Theory predicts the fall in the affected blocks, but does not predict the complete absence of displacement, unless the crimes that the potential thieves were deterred from committing were so close to the margin that walking an extra block or two to find another equally attractive target was not worth the effort, or the potential thieves had ties to the particular blocks they were operating on that could not be carried over to other blocks. Klick and Tabarrok (2005) for Washington, D.C., Poutvaara and Priks (2009) for Stockholm, and Draca et al. (2011) for London find similar results.

The types of crimes affected by simple presence are intuitive: auto theft and thefts from cars in Washington, auto theft in Buenos Aires, football and ice hockey hooliganism in Stockholm, thefts and street-level violence in London (burglaries respond to police presence in Washington, but not in London). In all these articles the implied elasticity of total crime with respect to police presence is about 0.3–0.5. Notice that this is an elasticity with respect to police, not with respect to the probability of punishment, the variable in Becker's theory. The closer one gets to the elasticities in Becker's theory, the less precise the estimates become.

"Hot spots"

Some types of crime, such as drug dealing and shooting, are highly concentrated in very small areas of cities. "Hot-spots" tactics flood these zones with visible police. A number of

¹³ As noted in Section 23.1, we do not discuss the causes or consequences of terrorism itself. For the sake of completeness, however, we note that there is a literature on the consequences of terrorism for urban form, ranging from the theoretical contributions of Mills (2002) and Glaeser and Shapiro (2002) to the empirical work of Blomberg and Sheppard (2007), Glaeser (2007), Abadie and Dermisi (2008), and Arbel et al. (2010). Researchers have also looked at how terrorism affects urban public finance (Wildasin, 2002) and residential segregation (Gautier et al., 2009).

experiments and quasi-experiments of hot-spots policing have been conducted and generally have found that it reduces crime in the targeted zones. No evidence of displacement has been found; reducing crime in hot spots does not appear to raise it elsewhere. Braga (2008) reviews these studies.

Police try many different approaches to crime, and even if none of them work, a few of them will generate statistically significant positive results. But the repeated results on terrorist responses and hot spots suggest that the positive results are not just reflecting good luck.

Perceived probability of arrest

Lochner (2007) finds that young men who say they have a higher probability of arrest if they commit a particular crime are less likely to commit that crime. Perceived probabilities are influenced by the ratio of crimes to arrests in the jurisdiction in which the young men live. Whether the reluctance of these individuals to commit these crimes leads others to do so instead is not known.

Notice that the results about police presence are also about the perceived probability of arrest. The theory is that a potential offender who is, say, contemplating stealing a car notices a police officer standing next to the car and thinks, “If I steal this car now, I will surely be arrested, probably before I get the door open.” That potential offender does not think, “Ten percent of the motor vehicle thefts around here are solved by arrests and that is the probability I will be arrested.” Consequently, there is no arrest and the arrest rate is in this case unaffected and irrelevant.

This example should also remind us that “perceived probability of arrest” is not a well-defined quantity; even for the same person it varies greatly with the circumstances under which the contemplated crime would occur. Operationally, the variable that Lochner uses is based on answers to a questionnaire that does not fully specify the circumstances of the contemplated crime.

The relevant magnitude for deterrence, moreover, is not the perceived probability of arrest, but the perceived difference between the probability of arrest contingent on offending, and the probability of arrest contingent on not offending.¹⁴

Actual arrest rates

Actual arrest rates—the ratio of arrests to reported crimes—are negatively correlated with reported crime, either contemporaneously or slightly in the future. This correlation was one of the strongest results of twentieth century criminology (Nagin, 1998). Four separate problems, however, make it hard to accept this correlation as causal, or the

¹⁴ This holds when a single decision is being made. O’Flaherty (1998) shows that in a repeated decision context, offending should be more sensitive to the false arrest probability than to the true arrest probability.

implied elasticity as an estimate of the Becker elasticity of crime with respect to the likelihood of punishment.

The first problem is measurement error. In the usual regressions, reported crime is both the dependent variable and the denominator of the independent variable, but it is only a noisy measure of actual crime (the goal of policing is to reduce actual crime, not reported crime, even if the incentives of officers are not perfectly aligned with this goal). This form of measurement error makes the absolute value of the elasticity estimated with reported crime greater in expected value than the elasticity that would be estimated with actual crime. When the reporting rate is high, the number of reported crimes is high and the arrest rate is low (the denominator is large), and this introduces a bias in the estimated elasticity, making it larger in absolute value. [Levitt \(1998a\)](#), however, argues that the bias from measurement error is probably small (the method is to show that estimates across a variety of equations do not differ in the way predicted by theory if the bias were sizable).

The second problem is endogeneity. The arrest rate is endogenous: for instance, extraneous reasons could drive up crime, overwhelm the police, and so drive the arrest rate down. Few articles attempt to deal with endogeneity. The main exception is the study of [Sampson and Cohen \(1988\)](#), who use a measure of police aggressiveness and professionalism as an instrument for arrest rates. They find a significant negative effect of arrest rates on robbery, but not on burglary. The exclusion restriction is that police aggressiveness and professionalism affects robbery only through the arrest rate. As the terrorism articles show, this restriction is dubious.

The third problem is that because the periods over which crimes and arrest rates are measured is finite, arrests can reduce crime from incapacitation as well as from deterrence. Suppose data are reported for a calendar year. Criminals arrested in January for crimes in January may spend all or most of the year in jail or prison if they cannot make bail or if they are tried and convicted within a few months. Arrest rates are also serially correlated, and so a high arrest rate in the current year is likely to be correlated with many previous year criminals incarcerated.

Levitt tries to estimate the incapacitation effect by looking at how arrests for one crime affect the committing of other crimes. Arrests for murder, for instance, should affect burglary only through an incapacitation effect; the murder arrest rate should not deter burglaries. Proceeding from this insight, he concludes that the incapacitation effect is important for some violent crimes such as rape, but not for property crimes or other violent crimes. He does not use an instrument for arrest rates though.

The final problem with arrest rates studies is conceptual. The probabilities of arrest that matter for criminal decision making vary by person, time, and circumstance. The arrest rate that researchers use is an average of these specific arrest probabilities. But it is not a simple average; it is an average over the opportunities when crime actually occurred.

To see what this means, suppose there are a lot of criminal opportunities (combinations of person, time, and circumstance) indexed by j and at each criminal opportunity there is a probability p_j that the perpetrator will be arrested if the crime is committed. There are n opportunities, each equally probable. In each of them the private benefit of committing the crime is the same, and the punishment if arrested is the same. Let $\hat{p}$ denote the maximal arrest probability for which committing a crime is profitable. The crimes that are committed are the set A or “active opportunities”:

$$A = \{j \mid p_j \leq \hat{p}\}.$$

Then the observed arrest rate is

$$\frac{1}{|A|} \sum_{j \in A} p_j.$$

Police tactics alter the vector $(p_1, \dots, p_n)$. The arrest rate acts like the Becker probability of punishment only if police tactics always cause uniform percentage shifts in p_j for $j \in A$, and do not change the membership of A . This is a strong restriction, and is unlikely to reflect good police work.

Suppose, for instance, that police are concerned about reducing crime, and have been given a small amount of additional resources. They would want to concentrate the new resources, to the extent they could, on those opportunities j for which p_j is slightly less than $\hat{p}$, and increase the probability of arrest there. This would reduce crime by reducing the membership of set A , but would leave the arrest probabilities for the remaining criminal opportunities that were still being acted on unchanged. But because the criminal opportunities that ceased being acted on were those with the highest arrest probabilities, the arrest rate falls—only the hard cases are left. So variation in the arrest rate will not reflect variation in crime or in the relevant arrest probabilities.

The arrest rate is therefore a poor proxy for certainty of punishment in Becker’s theory.

Probation sanctions

Probationers are people who have been found guilty of crimes, but sentenced to periods of supervision and restriction in the community, rather than to incarceration. They lose some of their freedom, but not all of it. If they are found to have violated the terms of their probation, they can be incarcerated or subjected to other sanctions. Project Hope, in Hawaii, modified traditional probation terms to include frequent drug testing and short (1- or 2-day) periods of incarceration in the event of drug test failure. Regular probation had less frequent testing, sanctions that were not always administered, and much severer punishments. So Project Hope increased certainty and reduced severity. In an RCT, the result was better compliance and fewer days in jail (Kleiman, 2009; Durlauf and Nagin, 2011).

Project Hope provides information on both certainty and severity. But since reducing the severity of the sanction was unlikely to increase compliance, it is a fairly convincing demonstration that certainty can increase compliance.

Number of police officers

Whether increases in the size of police forces are likely to reduce index crime is an important policy question, especially for state and federal governments that can provide funding, but may have little direct control over what police do. Since most (but not all) of what police do about index crime affects the certainty of punishment, not severity or incapacitation, it would probably be a weak confirmation of the proposition that punishment certainty reduces crime if police reduced crime. But a positive result about police force size is not necessary for the proposition on punishment certainty to hold: marginal additions to police forces could be frittered away on waste, sloth, or unproductive activity, or could be devoted to victimless crimes or traffic regulation.

The evidence seems quite clear now that, on average, increasing the size of a police force reduces index crime. Ordinary least squares regression is insufficient for this question because of endogeneity (places with more crime tend to have more police) and measurement error (the number of police is not measured accurately). Several different approaches to deal with this problem have been taken.

The commonest approach is to use instrumental variables to handle endogeneity, but without correction for measurement error. [Levitt \(2002\)](#) uses firefighter and other civil service employment as his instrument, [Evans and Owens \(2007\)](#) use federal subsidies, and [Lin \(2009\)](#) uses state sales taxes.¹⁵ All these articles find that police reduce crime, and that the effect is larger for violent crime, especially murder, than for property crime. The elasticities of violent crime with respect to police employment are 0.5 or more (sometimes greater than 1.0).

[Chalfin and McCrary \(2013\)](#) present evidence that police force size responds primarily to idiosyncratic shocks, and so endogeneity is not a large problem. But measurement error in their view is a substantial problem. They correct for measurement error, and find results similar to those in the other instrumental variables articles. Still, the murder elasticity is greater than 0.5.

The unresolved question is whether the entire crime reduction that police cause is due to deterrence. Arrests for index crime, as we have noted, often cause incapacitation. Police also pick up people on outstanding warrants, and sometimes when they arrest parolees or probationers for minor crimes, those arrestees are sent to prison (even if they do not appear as index crime arrests).

¹⁵ The 2002 note was Levitt's second article on this question. The first, [Levitt \(1997\)](#), used electoral cycles as an instrument. [McCrary \(2002\)](#) found weaknesses in this article, and Levitt responded with the new instrument and new results.

Chalfin and McCrary believe that the greater police impact on violent crime than on property crime is evidence for incapacitation, but the issue remains open. Some violent crime, such as robbery and bar-room brawling, necessarily occurs in public, where police can be found, while much property crime can be done more furtively—by definition, the perpetrator of a violent crime must confront the victim. The protective responses of potential victims may also matter; see [Section 23.4](#). The effect of police on crime measured here is the net effect after potential victims have adjusted their precautionary behavior.

23.3.2.2 Severity studies

The empirical literature on severity is much sparser than that on certainty, and the results are less definitive. The articles that show any effects are few. That may be because severity is hard to measure, or because severity is too heavily intertwined with incapacitation, or because it really has little or no effect.

Collective pardon

Perhaps the cleanest test of whether sentence severity reduces crime comes from a 2006 prison release in Italy. Because of overcrowding, Italian authorities declared a collective pardon and released large numbers of prisoners before their sentences ended. If they reoffended, their remaining sentences would be added to the sentence for the new crime. Thus, former prisoners who had longer remaining sentences faced severer punishment for future crimes. [Drago et al. \(2009\)](#) found that former prisoners facing severer punishment were less likely to reoffend. Their estimates imply an elasticity of recidivism with respect to sentence length of -0.74 , but the differences in question are on the order of magnitude of a few months.

Prison conditions

[Katz et al. \(2003\)](#) find that worse prison conditions deter crime. They use the death rate of prisoners as a proxy for prison conditions. The rates of homicide, violent crime, and property crime all go down when prison death rates go up.

Age of majority

Punishment is severer for adults than for juveniles, and so when individuals become 18 years old, they should reduce criminal offending. [Lee and McCrary \(2009\)](#) use individual-level data to look for a discontinuity at the 18th birthday. They do not find one (the point estimate is negative, but tiny and insignificant). [Levitt \(1998b\)](#) found a drop, but his data were cruder.

California sentence enhancement

California's Three Strikes and You're Out law took effect in March 1994. It mandated a 25-year minimum sentence for conviction on a third "strikeable" offense. Thus, sentence severity increased considerably in March 1994, and it increased more for people with more strikeable convictions.

Crime did not decrease noticeably in California in March 1994, but individuals with two strikes reduced offending (or became less likely to be convicted). [Zimring et al. \(2001\)](#) found a discontinuity in offending by this group in March 1994. [Helland and Tabarrok \(2007\)](#) compared future offending of people who were convicted of two strikeable offenses, and those who were "almost" convicted of a second strikeable offense: they were charged with one, but ultimately convicted of something else. Those with the second strikeable conviction committed less subsequent crime.

23.3.2.3 Violence reduction initiatives

Beginning with Operation Ceasefire in Boston, a number of cities have experimented with targeted programs to deter violence among high-risk individuals. These programs raise severity, certainty, and salience; they are surely designed to deter potential offenders, but the elements of deterrence are too intertwined to disentangle. The details differ, but most programs revolve around "call-ins." A dozen or two selected people—parolees or known gang members, for instance—are called to a meeting with law enforcement and community leaders. The law enforcement leaders tell the participants that they are being watched, and that the first gang or group to engage in violence will be severely punished. If guns are involved, federal rather than state imprisonment will follow. They show slides about recent long sentences for people the participants might know. Former federal prisoners talk about what federal prison is like: it's cold, it's far away, nobody visits, and you and your friends are a small minority amidst prisoners from all over the country.

Then community leaders and representatives of social service agencies promise to help anyone who wants to change his/her lifestyle, and let those people do so in a less public way. Families of murder victims talk about the pain they bear and urge the participants not to get themselves killed.

Thus, the violence reduction initiatives combine greater certainty of punishment (promised credibly since it is "the first" who are punished, not everyone everywhere), greater severity of punishment (worse prison conditions, not just more years), and enhanced incentives to abjure violence.

These initiatives have been studied in Boston ([Kennedy et al., 2001](#)), Richmond ([Raphael and Ludwig, 2003](#)), and Chicago ([Papachristou et al., 2007](#)). Generally, they appear to have reduced murder rates, although whether the early studies were large enough to justify their strong conclusions is open to question ([Raphael and Ludwig, 2003](#); [Cook and Ludwig, 2006](#)). The programs are evolving, however, and more data are accumulating.

Quite apart from their effects on certainty and severity, such violence reduction programs can also operate through a *coordinated* change in beliefs. A generalized expectation that others are less likely to kill can reduce the incentives for any given individual to kill preemptively—for instance, during an escalating dispute (O’Flaherty and Sethi, 2010c). We explore this mechanism further when discussing homicide in [Section 23.4](#).

23.3.2.4 What empirical studies tell us

Deterrence works. Some things—though not everything—that authorities can do will usually reduce crime. Finding empirical correlates of Becker’s certainty and severity elasticities, however, is probably impossible.¹⁶ Nevertheless, cost-benefit calculations for various strategies can be made.

23.3.3 Income, earnings, and employment

The young men who commit most index crimes are generally believed to be disproportionately poor, although good data on the legitimate income of arrestees is hard to come by. Possibly this relationship is explained directly by long-run criminogenic characteristics: the same characteristics (e.g., being a high school dropout, being a child of a lone parent, being impulsive, or having a poorly functioning prefrontal cortex) that incline people to commit index crimes may also incline them to earn little or no money in the legitimate labor market because these characteristics are unattractive to employers. Labor market conditions might have no independent effect on crime.

Alternatively, committing index crimes might be a time-intensive activity, and so people whose opportunities were most limited in legitimate labor markets (for whatever reason) would have a comparative advantage in committing index crimes. Poor young men might commit crime because of their incentives, not their characteristics. If incentives were the reason, moreover, the volume of crime would change as incentives changed.

The standard theory gives several reasons why labor market opportunities would affect the incentives to commit crime. First, planning and committing crimes takes time, and so does evading capture after a crime has been committed. People with lower market wages or none at all have lower opportunity costs.¹⁷ Second, being arrested and tried also uses time, some of which may disrupt one’s legitimate employment. The opportunity costs here include not only time, but the possible loss of one’s job. Legal representation may also be subsidized for people with low incomes. Third, incarceration uses a great deal

¹⁶ The exception to this is laboratory experiments, such as the experiment of Harbaugh et al. (2013). That experiment indicated that both certainty and severity reduce crime, and that subjects tested were generally risk averse. But since these variables cannot be recreated outside the laboratory, it is unclear how the estimates generalize.

¹⁷ Freeman (1996), however, provides some evidence that most crime does not absorb an extravagant amount of time, and one can easily maintain a full-time job while being an active criminal.

of time, and is obviously cheaper for people with lower expected wages and employment rates. Fourth, the stigma costs of a criminal conviction, especially the labor market losses, are also likely greater for people with higher earnings. A criminal conviction is likely to cost a college president more than a laborer.¹⁸

On the whole, it seems likely that not only monetary but also subjective welfare costs of committing crimes are greater for more affluent individuals. What about the benefits? Given standard assumptions about the marginal utility of income, a given monetary gain results in lower welfare gains at higher incomes. On this reasoning, property crimes and robbery, where the motive is acquisition, should yield greater subjective welfare benefits to the poor. For assault, murder, and rape, where the motives are more various, no definitive statement is possible.¹⁹

One weakness of this labor market story is that it seems to apply only to people attached to the labor market, while much crime is committed by teenagers whose attachment to the labor market is limited, whether they are rich or poor. But most of the arguments carry over fairly easily, since the costs are in the future. Opportunity costs of teenagers who are on a trajectory for well-paying jobs are higher in terms of incarceration and stigma, at least. School activities can be especially valuable to these teenagers, certainly more so than to those who do not expect to have attractive labor market opportunities later in life.

So it is plausible that people with worse labor market prospects should be more likely to commit index crimes, and that labor market conditions should affect crime by affecting incentives. Very little modern work has been done on the first implication. The most sophisticated studies ask how various income support programs affect recidivism by recently released prisoners. These studies find an income effect (not just a substitution effect), especially for property crime. [Berk and Rauma \(1983\)](#), for instance, in one of the earliest regression-discontinuity applications in any field, find that unemployment compensation reduces recidivism, even though it probably discourages work.

The second implication has been more extensively studied. In their survey of ecological regressions, [Pratt and Cullen \(2005\)](#) found that unemployment with length considered was the second strongest predictor of aggregate crime rates, out of 31. Poverty ranked 10th. But measures of unemployment that did not account for length were only moderately strong predictors.

¹⁸ Note, however, that certain costs associated with a conviction may be borne only by people who are poor: losing access to public housing or subsidized student loans is not a great hardship for the affluent. Furthermore, the ability to mount a credible defense requires wealth, so the expected likelihood of incarceration is almost certainly lower for those with greater wealth.

¹⁹ We are considering only index crimes here, but more generally, there are crimes that are expensive to commit and therefore possible only for those with sufficient wealth. It is hard, after all, to borrow money to hire lawyers to harass people you do not like, however much you may seek this end, but you do not need to borrow if you have wealth.

Simple regressions of crime on unemployment or wages face several problems. First, general measures of labor market conditions may be poor proxies for the labor market conditions that people on the margin of committing crime face. Second, because many of the costs of crime are in the future, short-run fluctuations, especially in unemployment, may not be very important, especially for teenagers. Third, business-cycle fluctuations may reduce the benefits of crime at the same time that they reduce its costs. There may, for instance, be less to steal when unemployment rates go up, and murder rates may fall because the drug trade shrinks. These three problems bias down the simple estimates of labor market incentive effects.

Three other problems bias these estimates up. Private and public protection efforts may be cut back when there is an economic downturn: cities hire fewer police, and private employers hire fewer security guards. Second, crime might reduce employment, especially in the cross section, if employers find a crime-ridden environment unproductive, or workers demand compensating differentials. And finally, both the fluctuations in crime and the fluctuations in labor market outcomes might be driven by fluctuations in characteristics (especially characteristics unobservable in standard data sets such as ADHD and low pulse rate), and so incentives might be playing no role.

Several articles have handled a substantial number of these problems and find that labor markets do affect crime through incentives. [Raphael and Winter-Ebmer \(2001\)](#) look at unemployment. They use many control variables to reflect the behavior of public authorities and potential victims, and specify variables that describe marginal offenders (e.g., they distinguish between male and female unemployment rates for rape). They also use oil price and defense contract shocks as instruments for state unemployment rates. Instrumenting resolves both the reverse causality problem and the difficulties with unobserved criminogenic characteristics. A higher level of unemployment raises property crime, but the effect on violent crime is mixed and small. It may even reduce murder rates.

[Gould et al. \(2002\)](#) look at both the wage rates of non-college-educated men and their unemployment rates. To correct for observable characteristics, they use state-year residuals from a wage equation rather than simply wages. They also use instrumental variables reflecting labor demand. For unemployment rates, their results are similar to those of Raphael and Winter-Ebmer. The wage rate effects are somewhat stronger than the unemployment rate effects, as might be expected, and often show sizable impacts on violent crime as well as property crime.

[Johnson et al. \(2007\)](#) show that at least part of the effect that these articles find is an income effect not a substitution effect. They look at federal relief aid in the Great Depression, and find that increases in federal aid to a city reduced property crime (including robbery). Work relief—assistance tied to required employment—reduced property crime more than unconditional aid did.

[Corman et al. \(2013\)](#), on the other hand, show an independent substitution effect. Welfare reform in the 1990s caused little change in the income of single-parent families,

but increased employment. They find that welfare reform reduced women's arrests for shoplifting. Thus, [Johnson et al. \(2007\)](#) show that there is an income effect, [Corman et al. \(2013\)](#) show that there is a substitution effect, and both articles suggest that the total effect is the sum.

It is a legitimate inference from these articles that part of the reason why poorer people commit more crime is that their incentives are different. But no study has asked how much of the income gradient in crime offending is due to labor market incentives. By the same reasoning, many of the effects we attributed to criminogenic characteristics in [Section 23.2](#) were probably overstated, since many criminogenic characteristics lead to lower wages and higher unemployment. The big gap is that no one has estimated a cross-sectional income or wage gradient in offending.

23.3.4 Race and incentives

Minorities, especially African Americans, are more likely to commit most index crimes, and, unconditionally, are more likely to be punished for committing index crimes.²⁰

Because the elasticities of crime with respect to both certainty and severity are less than 1, certain explanations for these two statements can be ruled out, at least for traditional punishments. It cannot be the case that minorities commit more crimes because they are punished *less* conditional on offending; this would not be consistent with the fact that they are punished more unconditionally. For instance, if the expected punishment for African Americans were less than that for non-Hispanic whites, the former would commit more crimes, but not enough that their arrests and presence in prisons would be greater. So the reasons for high minority involvement in crime must lie elsewhere.

The older consensus in criminology was that arrests for index crimes by race were roughly proportional to offenses and that convictions and sentencing for index crimes were not seriously biased ([Sampson and Lauritsen, 1997](#)). More recent work on prosecutions and sentencing has found many exceptions to this conclusion, but no consistent pattern, although often studies find that black and Hispanic defendants are treated more punitively. But some studies find the opposite ([Kutateladze et al., 2012](#)). Bail setting appears to be an exception, as [Ayres \(2001\)](#) found more restrictive conditions for black defendants, and [Kutateladze et al. \(2012\)](#) report on more recent studies with similar findings for both black and Hispanic defendants. Bail, though, affects the jail population, not the prison population or arrest rates.

²⁰ Both statements need to be qualified. Nobody truly knows who commits index crimes, but considerable evidence indicates that minorities commit most of these crimes disproportionately; see [Sampson and Lauritsen \(1997\)](#). Second, Hispanic involvement in the criminal justice system is poorly measured because many law enforcement agencies do not gather ethnicity information in ways that are consistent with federal government guidelines.

Minorities might be less effectively deterred if they *perceived* lower expected punishment, even if actual punishment were the same or greater. [Lochner \(2007\)](#) reports results on perceived punishment, but these do not paint a consistent picture. In the National Longitudinal Survey of Youth 1997 cohort, blacks and Hispanics perceive lower probabilities of arrest for property crimes, both unconditionally and conditional on the official arrest rate in their county, and no information is available on violent crime. In the National Youth Survey, blacks and Hispanics perceive higher arrest probabilities for petty property crimes, but essentially the same arrest probabilities for violent crimes and more serious property crime.

On perceived severity of sanctions, we are not aware of any conclusive findings. On the one hand, minorities might be more comfortable (less uncomfortable) in prisons dominated by minorities; they might feel less stigma in their communities because their communities contain more ex-offenders; they may not lose as much income from employer disfavor because they would not have had as much income to start with. On the other hand, losing voting rights and dignity may sting more deeply for blacks because of the long struggle it took to achieve them in America; loss of housing subsidies may hurt more because they would have relied on them more; and employers may be more willing to look at a white ex-offender as a “kid who made a mistake” rather than a hardened criminal. In fact, in an employment audit, [Pager et al. \(2009\)](#) find that a criminal record is a greater penalty for black men than for white men. No answer about punishment size is obvious.

Repeated temptations may also matter, especially if minorities think they are more likely to be punished unjustly and are more likely to live in neighborhoods where temptations occur more often. Under such circumstances, some of the results of the Becker model of “one-shot temptations” do not carry over ([O’Flaherty, 1998](#)). Essentially, people who think they will go to prison eventually (because of wrongful conviction or inadvertent mistakes on their part), no matter what they do today, will not be much deterred by threats of punishment for today’s crime (especially if they do not discount the future heavily). But no empirical work has been done on this question as far as we are aware.

Earnings and employment are one area where racial differences are clear, and where evidence shows an incentive effect. Blacks and Hispanics earn lower wages than non-Hispanic whites when they work, and are more likely to be unemployed or out of the labor force (independent of any criminal justice system actions, apparently). Thus arrest and imprisonment should be less of a deterrent.

Putting a number on this effect, however, seems impossible at this time. Recall that no high-quality studies look at cross-sectional variation in offending by earnings. It is hard to see how the time series results could be translated into the cross section. Cross-sectional effects, for instance, are likely to be nonlinear, but the time series studies have understandably looked only for linear effects. A serious cross-sectional analysis would also have to model the earnings and employment outlooks of teenagers.

The time series studies found larger labor market effects for property crimes than for violent crimes, and practically no effect for murder. Racial differences in offending for property crimes are much smaller than those for murder and robbery. If the time series results carry over to the cross section, it seems likely that earnings and employment explain a much larger part of the racial gap for property crime than for robbery and murder. That is, the significant racial disparities in offending and victimization for homicide and robbery cannot easily be attributed to income and employment differences. We return to this issue in more detail when looking at interactions in [Section 23.4](#).

In summary, then, labor market incentives possibly explain a considerable part of the racial gap in property crime offending, but cannot account for the homicide and robbery data. We have no reason to believe that any particular racial or ethnic group is especially likely to be deterred from index crimes by existing punishments. Arguments exist to suggest all manner of effects, and many questions still lack answers.

23.4. INTERACTIONS

While Becker's concern was with the optimal design of monitoring and enforcement systems, his general approach provided a versatile framework within which a variety of related phenomena could be examined. [Cook \(1986\)](#) considered interactions between potential victims and would-be offenders, with the former attempting to protect themselves and the latter seeking to find the most lucrative targets. Victim selection, from this perspective, is seen as an economic activity in its own right and responsive to incentives. Potential offenders will tend to target those who are easy to find, offer minimal resistance, have a lot to offer, and have limited recourse to the law. But victims who are the most attractive targets will also be the most inclined to use defensive measures: avoiding areas with high victimization rates, moving in groups, installing alarm systems, and so on.

The net result could well be lower victimization rates per capita for those who would otherwise be the most desirable targets. This helps account for the fact that elderly women are less likely to be robbery victims than young men, and banks are held up less often than gasoline stations. A high level of fear among some groups can result in so great a decrease in exposure to crime that even if victimization is high per unit of exposure, measured aggregate victimization rates will be low for the group.²¹

This perspective also suggests that the removal of some individuals from the population of potential offenders through incarceration or rehabilitation need not significantly reduce overall crime rates. Faced with less competition, the remaining offenders would

²¹ The importance of precautions in determining rates of victimization may be illustrated by considering fatalities from lightning strikes: over the period 2006–2012, more than 80% of such victims were male ([Rice, 2014](#)). While some of this may be attributed to physiological gender differences such as body mass, it seems likely that the main cause is differential exposure to conditions in which one is vulnerable to a strike.

see an increase in returns. This could cause them to scale up their activities, and could bring in new entrants that swell their ranks. That is why we had to distinguish between strong and weak claims of criminogenic characteristics.

Interaction effects are also important within the population of potential offenders. Those engaged in activities such as drug selling, gambling, or prostitution are often attractive targets for robbery, both because they are likely to be in possession of cash or other valuables, and because they are unlikely to report the crime. Street vice also requires location choices to be made, and these clearly involve interactions with the choices of others. Homicide is sometimes preemptive, or retaliatory, or motivated by reputational concerns, which calls for a strategic analysis. Effective law enforcement critically depends on witness cooperation, the costs of which depend on the behavior of other potential witnesses. In this section we deal with interaction effects such as these.

23.4.1 Private actions and displacement

Private actions, sometimes in concert with public initiatives, are a major factor in determining the aggregate level and spatial distribution of crime. In fact, without private action it is impossible to imagine any crime control policy being effective. For one thing, the willingness of victims and witnesses to report crimes, provide information to police, and testify in court are all crucial inputs in the production of law enforcement. The use of alarms and locks, hiring of security guards, installation of cameras, and use of credit and debit cards as substitutes for cash all affect the returns to property crime. Technological innovations that make it easier to trace stolen vehicles or uncover fraudulent credit card use all play a role. Electronic sensors on high-value items in stores make theft of merchandise harder to execute. Well before high-tech surveillance became possible, [Jacobs \(1961\)](#) celebrated the “eyes on the street” that deterred crime in Greenwich Village.

Private actions to avoid victimization can have significant external effects, both positive and negative. For instance, private security guards face an incentive to protect a particular area or premises, not reduce crime overall. This can result in displacement of crime to other venues, increasing inequality in exposure to crime without lowering the aggregate volume. However, if not all criminal opportunities are equivalent, the protection of the most lucrative targets will lower the overall returns to crime ([Ehrlich, 1973](#)). On the margin between criminal and noncriminal activities, therefore, the protection of some locations should raise the relative return to noncriminal activities and lower the volume of crime. For example, better protection of high-value motor vehicles through improved locking or tracking systems shifts crime to lower-value vehicles, while also lowering returns to motor vehicle theft, and gated communities shift burglary to communities without gates and lower the returns to burglary ([Helsley and Strange, 1999](#)). Displacement (and negative external effects on others) occurs together with an overall reduction in crime.

This point is very clearly illustrated in a study by [van Ours and Vollaard \(2013\)](#) on the electronic engine immobilizer, an inexpensive device that makes motor vehicle theft substantially more difficult and costly unless the offender has access to an original key. A European Union requirement made installation of the device mandatory for all new cars as of 1998, and the proportion of new cars equipped with the immobilizer rose steadily from essentially zero to 100% over the 5-year period starting in 1994. The study authors examine theft data from the Netherlands using the policy shift as an exogenous shock and find evidence of a substantial reduction in aggregate motor vehicle theft, sustained over a decade, and only partially offset by target displacement. Roughly one-third of the decline in the theft rate for cars with the device was shifted to older cars. Since the stock of cars without an immobilizer was steadily dwindling over time, this displacement effect quickly became negligible. The study authors conclude that the policy comfortably survives a cost-benefit analysis.²²

Ordinarily a mandate like this would be justified by a significant market failure, which in this case is hard to identify. The cost of the immobilizer is small relative to the *private* benefits to most car owners and insurers, even if one disregards any positive external effects from installation.²³ Furthermore, the regulation was a blunt instrument that took no account of differences in the risk of theft faced by individual owners. As the study authors note, a policy aimed at lowering “a risk that is already low through a one-size-fits-all measure that does not distinguish between targets at high or low risk sets a high threshold for achieving positive net benefits.” Nevertheless, given the limited market penetration of the technology in countries where it is not mandated, the study authors conclude that in decisions involving rare events, individuals do not always act in their own best interest.

It is important to note that, in principle, displacement in response to a deterrence policy can result in *increased* social costs relative to the status quo, even if one entirely disregards the cost of implementing the policy. For instance, the mandatory use of an immobilizer could have induced would-be car thieves to switch to carjacking instead, a crime with substantially greater victim costs (including the costs of injury). Fortunately, this appears not to have occurred on a scale that would lead one to question the conclusions of the study, but tactical displacement that sharply raises victim cost must be considered in any prospective evaluation of such measures.²⁴

²² Along similar lines, [Vollaard and van Ours \(2011\)](#) examine the effects of a Dutch law mandating burglar-resistant doors and windows in all new construction. They find that homes built after the mandate came into effect had significantly lower rates of burglary, with no evidence of displacement to burglary of older homes or to related property crimes.

²³ One such effect is that a lower theft rate reduces the deadweight loss associated with police activity. Some of these costs can be transferred to the owner in the form of fees for vehicle recovery and storage.

²⁴ For instance, a recent spike in carjackings in Newark has been attributed to the increasing use of engine immobilizers ([Santora and Schwartz, 2013](#)).

Helsley and Strange (2005) explore the interaction between private and public crime reduction initiatives, where only the latter can be used to implement increasing marginal penalties for crimes of greater severity. When the two types of protective activities (public and private) are strategic substitutes, an increase in private protection by some individuals or communities results not only in displacement to other venues, but also in a contraction in public enforcement, and an overall increase in the severity of crime. There is excessive use of private protection in equilibrium (relative to the first best), suggesting that private measures should be taxed, not subsidized or mandated, since they generate negative externalities.²⁵

Private crime prevention measures can also give rise to positive external effects. One of the most striking examples of this is LoJack, a concealed device capable of transmitting radio signals to receivers. Although the costs are paid on installation by the vehicle owner, the benefits spread widely because the infrastructure for motor vehicle theft can be disrupted. Ayres and Levitt (1998) found substantial positive external effects from the installation of LoJack, with benefits exceeding costs by an order of magnitude. Even a low rate of installation is enough to generate substantial benefits since the professional thieves with high turnover are unlikely to avoid contact with an equipped vehicle for long. A crucial part of the success of the device come from the fact that no visible signs of installation are present. A strong case could be made for subsidizing such general deterrence measures so that some of the external benefits are internalized. For instance, if insurance companies are required to lower rates for LoJack-equipped vehicles, they may all benefit, even though none have the incentive to unilaterally lower rates by much.

Cook and MacDonald (2011) survey the evidence of the role of private actions in crime deterrence, and present fresh evidence on the effectiveness of business improvement districts (BIDs) in Los Angeles. These are public-private partnerships, allowing levies to be placed on (even reluctant) business owners to finance security and other aspects of neighborhood enhancement. Cook and MacDonald find benefits far in excess of the costs, with significant declines in robbery, burglary, and motor vehicle theft rates. The reduction in crime is not generated by an increase in arrest rates, and therefore does not impose additional costs on public institutions. In fact, arrest rates are found to decline significantly. Furthermore, there is little evidence of displacement: if anything, there are modestly positive spillover effects to immediately neighboring areas.

23.4.2 Deterrence and selection

The idea that crime is an economic activity like any other, subject to entry and exit based on the usual considerations of benefit and cost, has some unexpected implications once

²⁵ An even broader range of negative externalities from private protection is considered by Helsley and Strange (1999), who show that the spread of secure residential communities can shift crime to less well protected business districts and thereby reduce aggregate employment opportunities. This, in turn, can increase the incidence of crime.

offender heterogeneity is taken into account. For instance, deterrence measures that are successful in reducing the volume of crime will do so by encouraging exit, but those who exit will not be randomly drawn from the pool of current offenders. Those who remain active despite greater deterrence will be precisely those with less attractive outside options. If this attribute is correlated with others, such as desperation or a propensity to violence, then the nature of crime will change along with its magnitude.

This idea is explored by O'Flaherty and Sethi (2009a) in the context of robbery. Robberies may or may not involve victim resistance and, conditional on resistance, they may or may not involve violent attempts by offenders to force compliance. If successful deterrence results in the selective exit of those offenders who are least likely to respond to resistance with violence, then the proportion of robberies resulting in victim injury will rise even as the overall incidence of robbery falls. This is indeed what we see in the data. Over the 12-year period 1993–2005, total robberies declined significantly according to National Crime Victimization Survey data, but robberies involving victim injury declined much less dramatically. Hence robberies became more violent even as they became less frequent. Furthermore, the likelihood of violence conditional on resistance rose sharply. There was some decline in the rates of resistance, as one would expect given the great likelihood that resistance would be met with violence, but on this point the data are less clear-cut.

23.4.3 Stereotypes

Victim resistance is itself a form of deterrence, and is taken into account by offenders as part of the process of victim selection. In their ethnographic study of active armed robbers, Wright and Decker (1997) provide some interesting anecdotal evidence on this point. While most of the offenders in their sample were black, many expressed a clear preference for white victims. This was driven by a belief that they would encounter less resistance from white victims, on account of exaggerated stereotypes about black male violence that are not shared by black victims.²⁶ Interestingly, a preference for white victims arises despite the fact that such victims are *less* likely to be carrying large amounts of cash.²⁷ It is clear that the likelihood of resistance rather than the anticipated take is the main criterion for victim selection. The lower perceived likelihood of resistance is itself

²⁶ One offender confesses: "I rob mostly whites. . . I usually don't have no problem [with resistance], none at all. [Whites] got this stereotype, this myth, that a black person with a gun or knife is like Idi Amin or Hussein. And [a] person [who believes] that will do anything [you say]." In contrast: "A black person will try to grab the gun out of your hand. They will make you shoot them if you have to." Another claims: "Whites accept the fact that they've been robbed. . . some blacks would rather die than give you they bucks and you damn near have to be killing [them] to get it" (Wright and Decker, 1997).

²⁷ One offender complains: "most white people have about two dollars on them, and credit cards, something like that." Another echoes this sentiment: "whites, they have credit cards and checkbooks on them. . . they get robbed, they cancel it." And another: "all they got is plastic and checks" (Wright and Decker, 1997).

understood by some offenders in the sample to arise from money being easier to come by for whites on average, and losses therefore being small relative to overall wealth.

Stereotypes of black male violence have a number of other implications. They make resistance less likely when the offender is black, and this increases the relative return to robbery.²⁸ As a result, rates of offending should exhibit greater racial disparities for crimes such as robbery, in which stereotypes are relevant, than for those such as burglary or theft, in which no face-to-face interaction between victim and offender occurs (O'Flaherty and Sethi, 2008). Put differently, the stereotype affects the relative returns to different categories of crime among offenders, as well as the relative attractiveness of criminal versus legal activities. As a consequence, racial disparities in arrest and incarceration should be expected even if the criminal justice system were entirely nondiscriminatory.

This reasoning can help account not only for racial disparities in offending, but also in the manner in which the perceived identities of offender and victim affect the way in which a crime plays out. Given lower resistance by white victims, the pool of offenders who selectively target whites will include those who most fear resistance. And if these individuals are less likely to force compliance when confronted with resistance then, other things being equal, resistance by a white victim should be *less* likely to result in violence than resistance by a black victim, if the race of the offender is held constant. This somewhat surprising prediction turns out to have empirical support (O'Flaherty and Sethi, 2008).

The broader point is that stereotypes can significantly affect interactions between victims and offenders and thereby give rise to differences across groups in the returns to crime. But this effect can arise only for crimes such as robbery that routinely involve face-to-face contact. Stereotypes cannot come into play when the identity of an offender is unknown to the victim, as is normally the case for burglary or motor vehicle theft. As a result, racial differences in rates of offending are liable to be greater for some crimes than others, and any analysis of such gaps needs to take account of the fact that crime is an aggregation of a range of highly disparate activities.

23.4.4 Segregation

Racial residential segregation is a central concern in urban economics. Despite a decline from the high watermark around 1970, black-white segregation remains pervasive, especially in the larger cities of the Northeast and Midwest.

There are three broad categories of explanation for this phenomenon in the literature: (1) segregation by race is an incidental effect of sorting by income and racial income disparities, (2) discrimination in housing and lending markets prevents access to some

²⁸ If whites in particular hold exaggerated stereotypes, they will resist black offenders at lower rates than black victims, making whites as a group more attractive to offenders. This could explain, in part, why robbery crosses racial boundaries far more frequently than homicide or rape.

neighborhoods for blacks, and (3) preferences over neighborhood racial composition combined with decentralized uncoordinated location choice result in endogenous separation.²⁹

The role of stereotypes in affecting rates of crime victimization provides an alternative, complementary explanation for racial segregation. If whites are more attractive robbery victims than otherwise identical blacks, they will exit high-crime neighborhoods at greater rates, even if they are unconcerned about neighborhood racial composition and do not benefit from favorable treatment in housing or lending markets. This does not require stereotypes to be based on anything other than the income distribution of the group to which one belongs. For instance, if lower-income individuals are known to resist at higher rates, and there are racial disparities in the distribution of income, then those who belong to a lower-income group will be selectively avoided by potential offenders. Hence, two otherwise identical individuals who belong to groups with different income distributions will face different victimization rates at any given location. As a result, conditional on own income, those who are visibly identifiable as belonging to more affluent groups will exit high-crime neighborhoods more readily, giving rise to greater levels of racial segregation than would otherwise be observed (O'Flaherty and Sethi, 2007).³⁰

This reasoning helps explain why those who are more attractive victims at any given location might have lower victimization rates in the aggregate. Furthermore, it suggests that racial disparities in victimization rates will be greatest at intermediate income levels, at which individuals are affluent enough to contemplate a move to a safer location but can be induced to stay by a somewhat lower victimization rate.

The idea that racial disparities in crime victimization at a given location can give rise to disparities in willingness to pay to move to a safer location has implications for group differences in savings rates. In the United States, black-white differences in wealth far exceed differences in income (Blau and Graham, 1990; Oliver and Shapiro, 1995), and it has been argued that a significant portion of this sustained disparity stems from

²⁹ The first of these factors has long been recognized to be relatively unimportant as an explanation, since substantial segregation exists even within income groups; see Massey and Denton (1987), Farley and Frey (1994), and Sethi and Somanathan (2009). On discrimination, the most compelling evidence comes from audit studies; see especially Yinger (1986). The literature on sorting is descended from Schelling (1971); see Sethi and Somanathan (2004) and Card et al. (2008) for relatively recent contributions.

³⁰ Verdier and Zenou (2004) propose an alternative theory of segregation in which stereotypes and crime play a central role. In their model, employers in a central business district set wages on the basis of employee race and location, with lower wages for those who live further away and those who are believed to belong to a group in which criminal offending is more common. Even with ex ante identical groups, beliefs about a higher incidence of offending among blacks can be self-fulfilling because those who are stereotyped in this way have less to gain from employment and choose to live at lower-rent locations distant from the business district. This lowers the opportunity costs of engaging in crime and results in greater rates of offending.

differences in rates of saving out of current income (Altonji and Doraszelski, 2005). If whites outbid blacks for housing in safer locations (on account of being selectively targeted where crime is high), they will exhibit lower levels of nonhousing consumption out of current income. Equivalently, they will exhibit higher levels of personal saving and wealth accumulation because housing is both a consumption good and an investment good.³¹

23.4.5 Peer effects and strategic complementarity

Glaeser et al. (1996) argue that the incidence of crime is too variable across time and space to be explained by fundamentals alone. They argue for the importance of local peer effects, arising from the ability of individuals to influence their social neighbors. This can add an additional source of variation across locations, so otherwise identical cities and neighborhoods could have very different crime rates. It can also add volatility to the time series associated with a given location, in excess of any volatility arising from fluctuations in fundamentals.

Schrag and Scotchmer (1997) also discuss a number of pathways through which a rise in offending by some portion of the population results in increased incentives for others to offend. They point out that the incentives to offend depend on the payoffs of criminal behavior relative to that of innocence, and that greater crime rates can increase the former while reducing the latter. For instance, with a fixed enforcement budget, the likelihood of arrest if guilty can decline even as the likelihood of arrest if innocent rises. Similar effects arise in jury deliberations: high crime rates overall can result in greater rates of error in the determination of guilt. These are instances of *strategic complementarity*—greater offending by some makes offending more appealing to others—and this commonly gives rise to multiple equilibrium crime rates for a given set of fundamentals.³²

In the simplest version of the model of Glaeser et al. (1996), individuals are arrayed in a (social) circle and are of three types. The first two types are not subject to social influence, and either always commit crime or never do so. The third type is subject to peer influence, and simply imitates the behavior of the neighbor on the left. Any given distribution of types in the population is consistent with multiple levels of crime depending

³¹ This effect would arise if, conditional on income, whites outbid blacks to live in more expensive neighborhoods for any reason at all. Cutler et al. (1999) argue that this has been a consistent pattern in residential choices since around 1970, and Sethi and Somanathan (2004) show that this outcome is predicted in a model of equilibrium sorting when individuals care about both the mean income and the racial composition of their neighborhoods. Charles et al. (2009) provide an alternative and complementary theory of the racial savings gap, based on the idea that individuals use conspicuous consumption to signal that they have high incomes. The incentives to do so are greatest for those belonging to lower-income groups.

³² Strategic complementarity can also be a feature of victim behavior. Helsley and Strange (1999) show how gated communities can “spread like a contagion” since their increased prevalence concentrates crime outside their boundaries, and strengthens incentives for more and more households to seek better protection.

on how individuals happen to be arrayed in social space. If all those who commit crime independently of peer effects happen to be adjacent, and this string of individuals is bordered on the right by an individual who is not subject to peer influence, then crime will be at the lowest level consistent with the type distribution. But if all those subject to peer influence are adjacent to each other, and this string is bordered on the left by one who commits crime regardless, the crime rate will be maximal. These are extreme possibilities, and a variety of intermediate cases can easily be imagined. Interaction effects introduce a nonfundamental source of variability across space and time, and this effect is strongest when a large proportion of the population is subject to peer influence.

Conversely, volatility in crime rates relative to fundamentals can be interpreted as a measure of the strength of peer effects for particular crimes. Glaeser et al. (1996) are interested in estimating the size of the nonfundamental variability, interpreted as the strength of social interactions, for various crimes. They find extremely high levels of social interaction for robbery, assault, burglary, auto theft, and larceny, but much smaller effects for murder, forcible rape, and arson.

In interpreting these findings, one must keep in mind that homicide, forcible rape, and arson are rare relative to the other crimes considered. For instance, there are over 4000 larcenies per 100,000 persons in the 1986 cross-city data used by the study authors but fewer than nine homicides. Most individuals never seriously contemplate committing murder, and we suspect that among potential homicide offenders, a very small number are certain to kill.

In fact, there is one mechanism unique to homicide among major crimes that would suggest high levels of strategic complementarity, and thus variability across time and space: preemption. Murder can be motivated by fear of being killed, and this fear is itself related to the background homicide rate. The logic of this argument was brilliantly described by Thomas Schelling in his 1960 classic *The Strategy of Conflict* as follows:³³

If I go downstairs to investigate a noise at night, with a gun in my hand, and find myself face to face with a burglar who has a gun in his hand, there is a danger of an outcome that neither of us desires. Even if he prefers to just leave quietly, and I wish him to, there is danger that he may think I want to shoot, and shoot first. Worse, there is danger that he may think that I think he wants to shoot. Or he may think that I think he thinks I want to shoot. And so on. "Self-Defense" is ambiguous, when one is only trying to preclude being shot in self-defense.

This effect helps account for the substantial variability in homicide rates across time and space. For instance, the homicide rate in Newark doubled over the 2000–2006 period, while the national rate remained essentially unchanged. Many of these killings were

³³ See Baliga and Sjöström (2004) for a formalization of this idea, and a demonstration that cheap talk can substantially mitigate the problem.

traced to escalating disputes between acquaintances often over seemingly trivial matters (O’Flaherty and Sethi, 2010c).³⁴ In an environment where killing is common, the possibility of escalation to homicide is salient, and the preemptive motive for killing looms large. But this itself contributes to a climate of fear and can cause murder rates to remain stubbornly high. Small reductions in homicide may be very hard to attain under such circumstances, but large reductions are not impossible. An effective policy to reduce homicide rates requires a coordinated shift in expectations, something akin to a truce in a war zone.³⁵

“Climate” and “environment” in this model are often spatial. As in war, location often signals the type of behavior that can be expected. Just as behavioral expectations in bars differ from those in libraries, neighborhoods with a high prevalence of crime generate different expectations from those in which crime has historically been infrequent.

This reasoning also helps account for the very high rates of homicide victimization and offending among young black males. The preemptive motive for killing is operational in the presence of fear, and this depends not only on the homicide rate in the local environment but also on the observable characteristics of individuals engaged in a dispute. Those who are feared are more likely to be killed preemptively, and this can induce them to also take preemptive action, in a cascading cycle (O’Flaherty and Sethi, 2010d). These effects are especially stark when both parties to a dispute have reason to fear. Ironically, a criminal justice system that treats offenders lightly if their victims are black will also induce those potential victims to become homicide offenders. Those who fear for their lives will be induced to kill, and those who are feared will be more likely to be killed preemptively.

Along similar lines, Grosjean (2014) shows that areas of the United States that were originally settled by Scots and Scots-Irish have higher rates of acquaintance murder among white males. She attributes this to a “herding culture” that these settlers brought with them, in which a reputation for toughness was needed to prevent theft of one’s livelihood. The culture of honor served as a substitute for legal dispute resolution mechanisms in the South, while having no such effect in the North, where formal institutions were stronger. Although she does not explicitly mention this, it seems plausible that preemptive killing was an amplifying factor in this environment.

³⁴ This pattern arises quite generally: “About half of all homicide victims are murdered by neither intimate family members nor total strangers, but rather by people with some kind of pre-existing relationship: friends, neighbors, casual acquaintances, workplace associates, associates in illegal activities, or members of their own or a rival gang” (Reiss and Roth, 1993, p. 78).

³⁵ Formally, the shift needs to be large enough to allow for coordination on a different equilibrium with lower rates of killing. As in Schrag and Scotchmer (1997) and Glaeser et al. (1996), this model can have multiple equilibria. More generally, models of strategic complementarity exhibit an amplified response to any shift in fundamentals—a multiplier effect—even when equilibrium is unique.

23.4.6 The castle doctrine and stand-your-ground laws

These considerations have implications for the effects of stand-your-ground laws, which allow individuals to legally kill others when under perceived threat, even if an opportunity to retreat to safety is available. The Castle Doctrine refers to the principle that one does not have an obligation to retreat in one's own home (or "castle"), and stand-your-ground laws extend the principle to other locations at which one has a legal right to be present. Florida's 2005 stand-your-ground law was one of the earliest examples of this, and more than 20 other states have enacted similar legislation.

By broadening the range of circumstances under which killing is treated as justifiable by the legal system, such laws make threatened individuals more dangerous and hence also more likely to be killed preemptively. The effect is strong enough to show up in empirical tests. [Cheng and Hoekstra \(2012\)](#) group states into those that adopted stand-your-ground laws and those that did not, and examine within-state changes in homicide rates and other outcomes across the two groups of states. They find that the laws fail to have the intended effect of deterring burglary, robbery, and aggravated assault. They do, however, have an unintended effect: they increase murder rates by an amount that is both quantitatively and statistically significant. These are not killings in self-defense, which would be classified as justifiable homicide and not murder. This is consistent with the preemption story.

[McClellan and Tekin \(2012\)](#) use a different data source (US Vital Statistics rather than the Uniform Crime Reports) and also find significant effects on homicide rates in states adopting these laws. But the effect arises only for whites, and is especially strong for white males.

23.4.7 Police stops

Consider a police force that wants to prevent crime by incapacitating the people most likely to commit crime shortly before they do so. (This is the forward-looking sort of goal we considered in [Section 23.2](#), not the backward-looking goal we considered in [Section 23.3](#).) It will try to predict who will commit crime, and use all the information at its disposal to do so.

For most crimes, the group of offenders is very far from being a representative sample of the population at large. As a result, a police force that was completely unconstrained in its use of diagnostic or preventive stops would have an incentive to engage in widespread profiling. That is, it would direct officers to use perceived racial and ethnic affiliation, in addition to such criteria as age and gender, in determining which individuals to detain.

Racial profiling by law enforcement agencies has attracted considerable attention since the 1990s, in the wake of some high-profile lawsuits.³⁶ The earliest of these were

³⁶ See *Wilkins et al. v. Maryland State Police* (1995), *Chavez v. Illinois State Police* (1998), and *Floyd v. City of New York* (2013) for prominent examples.

responses to motor vehicle stops and searches on major highways, which appeared to target blacks and Latinos at rates that were 2–3 times as large as their share in the population of drivers (United States General Accounting Office, 2000). More recently, stop-and-frisk operations on city streets have come under similar scrutiny. The New York City Police Department engaged in over 4 million stops and interrogations over the period 2002–2012, at an annual rate of more than half a million since 2006.³⁷ Blacks were over 9 times as likely as whites to be stopped (Fagan, 2010), and only 0.15% of stops—15 out of 10,000—resulted in the confiscation of firearms.

One question that has received attention in the literature is whether the striking disparity across groups in the rates of detention reflects optimizing behavior on the part of police, motivated by the maximization of arrest or contraband recovery rates, or whether it also reflects biased targeting of specific groups in excess of levels warranted by race-neutral criteria. In an influential article, Knowles et al. (2001) developed a model with the property that equilibrium behavior on the part of police and motorists, in the absence of any racial animus, would imply equality across groups in the rates of contraband discovery. By this logic, differences in “hit rates” across groups are evidence of racial bias as an independent factor affecting stops, with groups that are victims of bias having lower hit rates. Applying this test to data on vehicle stops on I-95 by Maryland State Police, Knowles et al. (2001) find no evidence of discrimination against black (relative to white) motorists, but do find that Latino drivers are victims of bias. Sanga (2009) extends this analysis to roads in Maryland other than I-95, and finds that on these roads Maryland State Police discriminate against both blacks and Latinos.

The hit rate test is simple and easily implemented, but not without shortcomings. The logic of optimization (for incapacitation) by police clearly implies equality in *marginal* hit rates across groups (Becker, 1957; Ayres, 2001). That is, the hit rates for motorists in each group whose other observable characteristics barely meet the threshold for a stop should be virtually identical, otherwise a shift toward a group with a higher hit rate would result in increased detection of criminal activity. But these marginal motorists (or pedestrians) are hard to identify from the data. The equality of *average* hit rates implied by the model of Knowles et al. (2001) is derived on the basis of more stringent assumptions. In particular, the study authors assume that observable characteristics are independent of whether or not an individual has actually chosen to carry contraband. There can be no outward signs of guilt conditional on being guilty. But this means that for any given set of characteristics (young black males for example, or female drivers in luxury cars), the set of vehicles searched and the set of vehicles not searched are equally likely to be carrying contraband. In equilibrium, each individual chooses a likelihood of criminal

³⁷ Summary statistics by year have been posed by the New York Civil Liberties Union at <http://www.nyclu.org/content/stop-and-frisk-data>. The policy is formally known as stop, question, and frisk.

activity that is calibrated to leave police indifferent between searching and not searching, given that individual's set of characteristics.

In addition, [Dharmapala and Ross \(2004\)](#) show that the hit rate result is very fragile. Minor perturbations of the underlying assumptions can show that the I-95 data are consistent with discrimination against minorities, or with no discrimination, or with reverse discrimination.

As noted by [Bjerk \(2007\)](#), this indifference property of equilibrium is both empirically unappealing and implies that the police stops are unlawful. In particular, they violate the requirement that there be "reasonable suspicion" based on "specific and articulable facts" at the individual level.³⁸ This is a more permissive standard than probable cause, but still requires suspicion at the individual level and thus operates as a constraint on police action. The constitution does not permit stops based solely on characteristics that are unaffected by the decision to break the law.³⁹ But allowing for such outward signs makes the hit rate test invalid as a tool for the detection of bias.

Alternative models of profiling have been proposed by [Anwar and Fang \(2006\)](#) and [Bjerk \(2007\)](#), who allow for signals that can be used as a basis for reasonable suspicion. Anwar and Fang assume that in addition to race, officers observe a noisy signal that is informative of guilt in the sense that, all other characteristics being held constant, higher values of the signal imply a greater likelihood that contraband will be discovered by a search. The signal distributions may differ across groups, which provides a rationale for statistical discrimination even in the absence of any racial bias. If the identity of the officer is held constant, the search costs may be sensitive to the identity of the individual detained; heterogeneity in these costs is interpreted as racial bias. As in [Knowles et al. \(2001\)](#), officers are assumed to value the successful detection of contraband, and maximize their payoffs net of search costs. This requires setting a signal threshold for each group such that a search occurs if and only if the signal of guilt exceeds the prescribed threshold value.⁴⁰

In this model, equality of average hit rates neither implies nor is implied by the absence of racial prejudice. What the model does predict is that in the absence of prejudice, the rank order of hit rates across groups should not depend on the group to which the trooper belongs. Using data from the Florida State Highway Patrol, including

³⁸ *Terry v. Ohio*, 392 U.S. 1 (1968).

³⁹ For instance, officers who detain pedestrians under the New York City Police Department stop-and-frisk program must fill out form UF-250, which requires them to indicate the basis for reasonable suspicion. This can include the possession of certain objects (such as a slim jim or pry bar) or actions indicative of casing victims or locations.

⁴⁰ The model of [Bjerk \(2007\)](#) has a similar structure, and also yields the conclusion that pure statistical discrimination does not imply equality in average hit rates. The two articles differ less on the theory than on the use to which it is put. Anwar and Fang conduct an empirical test, while Bjerk examines theoretically the manner in which enforcement of a color-blind policy is likely to affect the level of crime.

information on the groups to which individual troopers belong, Anwar and Fang are unable to reject the null hypothesis of the absence of prejudice. They recognize, however, that their test has low power and that their findings should accordingly be interpreted with caution. Furthermore, they find significant heterogeneity across groups of troopers in their search costs, and also report that an application of the hit rate test to their data would lead to the conclusion that troopers were biased against black and (especially) Latino motorists.

Despite the shortcomings noted above, the hit rate test is simple, versatile, and widely used. The New York City Police Department stop-and-frisk data reveal greater hit rates for whites relative to blacks and Latinos, both for possession of weapons and for possession of other contraband such as drugs. [Gelman et al. \(2007\)](#) have used this to argue that the policy is implemented in a discriminatory manner: a resource shift with increased detention of whites and reduced attention to blacks and Latinos should increase the recovery of contraband. This interpretation has been called into question by [Coviello and Persico \(2013\)](#), who argue that there is considerable heterogeneity across police precincts in hit rates, and this alone can give rise to disparities in aggregate data even if hit rates *within* precincts are identical across groups. For instance, if precincts with low hit rates (for whatever reason) were also those with the greatest incidence of black and Latino stops, the aggregate data could exhibit a disparity in arrest rates by group even if the data for an individual precinct did not do so. The study authors find that controlling for this factor by allowing for precinct-level fixed effects reverses the conclusions from the aggregate data: hit rates for whites are modestly lower relative to those for blacks and Latinos. They recognize, however, that this could be due to discriminatory allocation of resources across precincts.⁴¹

The question of whether police stops are motivated by pure statistical discrimination or also involve racial bias is important from a legal perspective, but even statistical discrimination that meets legal constraints can impose significant costs on innocents who belong to groups with high rates of offending. The distress experienced by an individual who believes that he/she has been subjected to a humiliating and invasive search on the basis of racial or ethnic markers must be taken into account in any welfare analysis. Since these costs could be substantially mitigated by a policy that is credibly color-blind, it is important to consider the question of whether and to what extent color-blind policing policies would result in greater crime.

[Bjerk \(2007\)](#) and [Harcourt \(2006\)](#) argue that the effect of statistical discrimination on crime is theoretically ambiguous: a shift to a color-blind policy, under certain conditions,

⁴¹ Furthermore, if the issue of a summons (rather than an arrest) is used as the action for which a hit rate is computed, then the hit rate for blacks and Latinos is found to be lower than that for whites, consistent with the aggregate data. One possible interpretation of this finding is that conditional on a violation being detected, there is a disparity in response, with whites being more likely to receive a summons rather than to be arrested.

could even reduce overall crime. This is because maximizing the likelihood of catching those who are engaged in criminal activity does not take into account deterrence effects that could affect the overall incidence of crime. Any shift in targeting will make crime more lucrative for groups on which pressure is eased, and less appealing to others. If the former are small as a proportion of the total population, or relatively unresponsive to incentives, while the latter are more numerous or more responsive to incentives, then the overall incidence of crime could drop. In addition, if the likelihood of detection is small, stops will have limited deterrence value no matter how they are distributed across target groups, and a transition to a color-blind policy will not have large effects on the incidence of crime.

The effect of statistical discrimination on the volume of crime is not just theoretically ambiguous, it is also empirically hard to identify. Isolating the effects of a single police practice is in general very challenging because a whole range of tactics are often implemented simultaneously as part of an overall strategy. Even if the strategy as a whole succeeds, identifying the separate effects of its constituent parts is often difficult. New York City experienced a dramatic decline in crime over the two decade period starting in 1990, while implementing a wide variety of police initiatives including the targeting of hot spots, dismantling of public drug markets, use of the CompStat system for mapping and statistical analysis, gun recovery programs, stop-and-frisk policies, and low tolerance for minor “quality-of-life” crimes. There was also a substantial increase in the size of the police force in the early part of the period. In a comprehensive analysis of the evidence, [Zimring \(2011\)](#) concludes that the targeting of hot spots and drug markets was highly effective, while CompStat, gun recovery programs, and increased police numbers were also probably important contributors to the decline in crime. He finds no evidence to support the claim that stop-and-frisk policies played a major role.⁴²

If true, this is an important point because statistical discrimination can have significant welfare effects. Bearing in mind that the overwhelming majority of stops result in no evidence of wrongdoing, we see that the impact on the innocent is an important consideration. Here it matters a great deal whether targeted people believe that they have been stopped on the basis of reasonable suspicion at the individual level, or simply because they happen to fit a particular demographic profile. If the only goal of the policy were to reduce the incidence of crime to the lowest level consistent with the resources at hand, then equalization of arrest rates across groups may be a suitable measure of success. However, if equalization of arrest rates requires the use of demographic information on

⁴² In addition, Zimring argues that the New York City crime decline cannot be attributed to “broken windows” policing because this policy was never really tried. In particular, he maintains that statutes against prostitution and gambling were not consistently enforced, and there were sharp declines in arrest rates for nondrug quality-of-life offenses. There is also limited support from other cities for the effectiveness of this strategy.

race and ethnicity by police, then a substantial welfare cost may be imposed on members of particular groups.

This cost must be accounted for in any welfare analysis of the policy. It is not the condition of being stopped that is most irksome to an individual in a heavily targeted group, it is the realization that he/she has been targeted for no reason other than his/her group membership. In his commentary on the killing of Trayvon Martin, [Blow \(2013\)](#) made the point as follows: “The idea of universal suspicion without individual evidence is what Americans find abhorrent and what black men in America must constantly fight. It is pervasive in policing policies—like stop-and-frisk, and in this case neighborhood watch—regardless of the collateral damage done to the majority of innocents.” No welfare analysis that fails to account for this collateral damage can claim to be complete.

Since the costs imposed on innocents are not internalized by police, the scale of profiling is likely to be inefficiently high. Direct compensation of the innocent is not feasible for several reasons. It would create incentives to act in ways that invite stops, thus eroding the informativeness of suspicious behavior. Corruption could be encouraged if friends are stopped as a favor. And unless compensation is contingent on race or ethnicity, it would not result in the internalization of identity-specific costs of public humiliation.⁴³

23.5. INCARCERATION

Incarcerating someone who has committed a serious crime, is likely to commit more crimes in the future, and who will commit fewer crimes after release because of the prison experience ostensibly serves many purposes. Retribution is gained because bad deeds have been punished. Deterrence occurs because punishment follows crime. Incapacitation results because prisoners usually cannot commit crimes against nonprisoners. And rehabilitation happens to the extent that beliefs and preferences are altered. Incarceration appears to be an all-purpose answer to many different questions.

But like a vacuum cleaner that also plays music and mixes fruit smoothies, or (to use a more familiar example) a jack-of-all-trades who is master of none, incarceration may not

⁴³ Some degree of cost internalization can arise through media exposure and legal recourse, both of which have played prominent roles in the debate over stop-and-frisk tactics. The importance of *voice* as a means of improving organizational performance was stressed by [Hirschman \(1970\)](#), and media pressure has had a major impact on the investigation and prosecution of a number of high-profile murder cases in India ([O’Flaherty and Sethi, 2009b](#)). But public pressure in response to especially heinous crimes can also lead to coerced confessions and wrongful convictions, as in the 1989 case of the Central Park Jogger, where five juvenile defendants served out lengthy terms before having their convictions vacated in 2002 ([McFadden and Saulny, 2002](#)).

do any of these four jobs well because there are inherent conflicts among them. The major conflict is between deterrence and incapacitation; deterrence is backward-looking and asks what crimes someone has committed, incapacitation is forward-looking and asks what crimes someone will commit. The optimal incarceration strategy for deterrence is unlikely to be the optimal incarceration strategy for incapacitation, and vice versa. This is the same issue that arose with police stops, and [Harcourt \(2006\)](#) presents a unified treatment.

Since incarceration deprives people of rights and freedoms they normally enjoy, it is supposed to be reserved for those who have already committed serious crimes; retribution and deterrence by law and tradition are supposed to trump incapacitation and rehabilitation, at least at the jailhouse door. But forward-looking considerations enter into parole, bail, probation, and sentencing decisions, often through the use of formal risk assessment tools; for a history and discussion, see [Harcourt \(2006\)](#), chapters 2 and 3).

In this section we will focus on incapacitation effects, since we have already discussed deterrence and economists rarely write about retribution. We will also say a little bit about rehabilitation. Because the goal of incapacitation is to alter the distribution of criminogenic characteristics in the nonincarcerated population, it is like the policies we discussed in [Section 23.2](#). Accordingly, we must ask both about the weak claim—do incarcerated people commit fewer crimes than otherwise identical people who are not incarcerated?—and the strong claim—does more incarceration cause less crime?

We will also be concerned about how mass incarceration affects the neighborhoods prisoners leave and return to.

23.5.1 Accounting for trends

The incarceration rate in the United States at the beginning of the twenty-first century was extraordinarily high in two respects: it was higher than it had ever been in US history, and it was far higher than the incarceration rates in other developed countries.⁴⁴ Males, African Americans, and, to a lesser extent, Hispanics are disproportionately incarcerated.⁴⁵

⁴⁴ Before the late 1970s, the rate was stable for many years at less than 200 per 100,000 ([Raphael and Stoll, 2009a](#), pp. 3–4). At the peak, in 2009, 756 of 100,000 residents in the United States were incarcerated in adult facilities, including federal and state prisons and local jails ([Glaze, 2010](#), Appendix, table 2). Among developed countries, only Russia, with a rate between 500 and 600 per 100,000, approaches the US rate; rich countries such as Great Britain (153), Norway (73), France (96), Germany (85), Japan (58), and Canada (117) all incarcerate far fewer people ([International Centre for Prison Studies, 2011](#), data for 2008–2011).

⁴⁵ In 2008, the incarceration rate for African American men overall was about 4640 per 100,000 and for Hispanic men it was about 1650 per 100,000. For African American men aged 30–34 years, the imprisonment rate was 8032 per 100,000, and the overall incarceration rate, including jails, was probably over 11,000 per 100,000.

The rise in incarceration since 1980 is often attributed primarily to drug enforcement, but this is not accurate in any accounting sense. The number of people incarcerated on drug charges rose from 41,000 in 1980 to about 539,000 in 2008—an enormous increase, but still only about 30% of the total rise in incarceration over this time period (the fraction of population-adjusted rise would be somewhat greater).⁴⁶

However, drug sentences are shorter than other sentences, and so in comparison with those in prison at any point in time, drug prisoners represent a higher proportion of those admitted and released from prison. For instance, about 45% of state prison admissions are for drug offenses, although only about one-fifth of state inmates are being held for these offenses (Raphael and Stoll, 2009b). When we look at the stock of former prisoners, drug incarceration weighs much more heavily than if we look at the stock of current inmates.⁴⁷

Why, then, did US incarceration grow so tremendously between 1980 and 2008? Raphael and Stoll (2009b) decompose the changes in the state prison population—the largest component of incarceration—between 1984 and 2002, using a model in which individuals can transition between four states: free, incarcerated after committing a crime, on parole, or incarcerated following parole revocation. In the steady state of their model, the odds of being in prison depend on the crime rate and the certainty and severity of punishment. Each of these variables depends partly on private actions and partly on government actions. The crime rate depends on potential offenders and potential victims, both private parties usually, but obviously governments can do many things to influence the crime rate. The certainty of punishment depends on government efforts to apprehend and convict criminals, but also on the efforts of offenders to avert detection and capture, as well as the willingness of victims and witnesses to help authorities. The average time served depends on the government's sentencing rules and guidelines, parole policies (for both granting and revoking parole), and efforts to enforce parole conditions, but also on the willingness of prisoners and parolees to abide by the rules imposed on them.⁴⁸

Raphael and Stoll (2009b) compare implicit steady states of their model; since half the prison population turns over in a year, they argue that steady states provide a reasonable approximation to a long-term process. Between 1984 and 2002, they find slight decreases

⁴⁶ Drug incarceration numbers for 1980 come from *The Sentencing Project* (2011). For 2008, the numbers of drug offenders in state and federal prisons come from the *Bureau of Justice Statistics*, 2008, tables 6.001 and 6.57, respectively. For jail inmates, we used the total number from Glaze (2010) and the 2004 proportion of drug offenders in the jail population from the *Bureau of Justice Statistics*, 2004, table 6.19.

⁴⁷ We are grateful to Steve Raphael for this observation.

⁴⁸ This model makes clear that there is no simple mapping from government policies to incapacitation, either in the aggregate or for groups. A harsh government might have a large prison system—because prisoners serve long sentences—or a small one—because everyone is afraid to commit crimes or no one is willing to cooperate and testify. A lax government could have many prisoners because many crimes are committed, or few prisoners because prison stays are short and few people are sent to prison.

in most crimes (per capita). The crime rate is clearly not the major reason why the state prison population grew.

Prison admissions per crime rose for assault and for rape, but were steady for most other crimes. Drug crime admissions per capita, however, rose fivefold, and a large portion of this increase (or all of it) was due to a rise in admissions per “crime.” How much, of course, depends on the definition and measurement of drug crimes.

The expected time served before parole rose by about one-third for most crimes, but more for larceny and rape. (Because admissions for less serious crimes rose more than admissions for more serious crimes, average sentences did not increase.) A major change in severity, however, came from parole. Between 1980 and 2003, the proportion of parolees returned to custody per year rose from 13% to 29%. The average time served after returning from parole rose about 13%.

With three factors for each crime (per-capita offending, prison entries per offence, and prison exit rate) and three categories of crime (index crimes, drug crimes, and parole violations), no decomposition of the change in the steady-state prison population is going to be unique or natural. Raphael and Stoll include drug crimes with the other crimes, and do a separate analysis of parole. Because we wanted to look at drugs separately, we undertook a different decomposition using their tables 2.2 and 2.3. This decomposition is shown in [Table 23.1](#).

Of the increase of 272 per 100,000 in the state prison population, roughly equal parts were due to more index crime offenders (mainly for rape, assault, and burglary), more drug offenders, and more parole violators, with index crime offenders being most numerous. The increase in simple time served is responsible for about half of the index crime increase, but very little of the drug or parole increases. (This was calculated by increasing the time served to the 2002 level, but holding admissions per crime and crimes per capita at their 1984 levels). [Raphael and Stoll \(2009b, chapter 4\)](#) find that the increase in time served was largely due to policy change, such as mandatory minima and truth-in-sentencing laws.

Table 23.1 Sources of growth of per capita state prison population, 1984–2002
Proportion of total increase in the steady-state population per capita

All sources	100.0%		
Index and other nondrug crimes		38.0%	
Greater expected time			17.8%
Other changes			20.2%
Drug crimes		28.8%	
Greater expected time			1.6%
Other changes			27.2%
Parole violators		33.2%	
Greater expected time			1.3%
More prisoners			10.4%
Other changes			21.5%

Sources: [Raphael and Stoll \(2009b\)](#) and authors' calculations. See the text.

The increase in prison admissions per crime was a somewhat larger driver of the increase in index crime prisoners. Prison admissions per crime could rise because of more effective police work, because probation was used less often, or because criminals worked in larger groups. [Raphael and Stoll \(2013, chapter 2\)](#) show that the increase in admissions per crime for index offenders (for a slightly different period) is due almost entirely to an increase in prison admissions per arrest. The change in crime by itself would have reduced the number of index crime prisoners; in an analysis not shown, we found that changing crime per capita to its 2002 level but keeping time served and admissions per crime at their 1984 levels would have decreased 1984 index crime prisoners per capita by about 13%.

Why were there more parole violators in prison? Part of the reason is more parolees, and part of the reason for more parolees is more prisoners. We made a crude attempt to correct for this (much less sophisticated than that made by Raphael and Stoll) by asking what the admission rate for parole violation would have been if it had increased at the same rate as the nonparole prison population. About one-third of the increase in parole violators is due to the expansion of the prison system; the majority is due to a greater rate of parole revocation.

So the detailed picture is complicated. The leading causes of the prison increase are more drug admissions, more admissions per arrest for index crimes, longer time served for index crimes, and a higher rate of parole revocation. Index crime went down, but the prison population grew along a number of other margins. But, the basic picture is simple; policy changes made the state prison system more punitive.

To our knowledge, no similar analysis has been published for jails. The story for the federal prison system is similar: it grew because of policy changes ([Raphael and Stoll, 2013](#)). The federal system, although smaller, grew faster than the state systems: between 1980 and 2010, the number of federal inmates per 100,000 population rose from 11 to 67. Most of this increase came from drug inmates: the proportion of federal prisoners being held on drug charges rose from 28% in 1974 to 55% in 2004. The proportion being held on weapons and immigration charges also rose, while the proportion held for property crimes and fraud fell.

The rate of admissions per year per 100,000 more or less doubled, and the average stay tripled, to produce a sixfold population increase. The rate of arrests for drug, weapons, and immigration grew substantially, and so did prison admissions per arrest. The growth in admissions reflects not just policy changes; the number of federal crimes increased, and many issues that state courts handled before moved to federal courts. The same sort of policy changes that raised state sentences also drove the increase in federal sentences. (The federal prison system does not have parole.)

23.5.2 Empirical evidence on incapacitation

Several different strands of research have tried to assess the effect of incarceration—primarily state imprisonment—on index crime. In [Section 23.3.2](#) we surveyed the

literature on deterrent effects. For incapacitation, separate strands look at the weak claim and the strong claim. A final group of articles dealing with the strong claim examine the combined effects of deterrence and incapacitation directly, without trying to separate the two stories.

Even though around 750,000 people are in jail on an average night, and each year around 12 million people spend at least one night in jail (Minton, 2012), very little is known about how jail affects crime or human capital.

23.5.2.1 The weak claim of incapacitation

The simplest studies ask prisoners how many crimes they used to commit each year before they went to prison. Inmate surveys of this type give answers that cluster around 16–20 index crimes per year of imprisonment (Bushway and Paternoster, 2009). The implicit counterfactual is that had they not been imprisoned, prisoners would have continued behaving in the way that they behaved in the last year before they were imprisoned. Studies using matched samples produce estimates about half this size.

The only natural experiment on this question was studied by Owens (2009). In Maryland in 2001, sentencing guidelines on the use of juvenile arrest information changed in such a way that sentences for certain 23–25-year-olds became shorter. Her estimate is that the average prisoner in the affected group would have committed 1.5 more index crimes per year.

Bushway and Paternoster (2009) point out that these disparate estimates are not necessarily inconsistent, since the studies were measuring different things. The inmate surveys date from before the large rise in incarceration and so plausibly prisoners in that era were more elite and dangerous.⁴⁹ The group that Owens studied, young prisoners for whom juvenile arrest information mattered, may also have been unrepresentative. The incapacitation effect of prison depends on who is being incapacitated.

23.5.2.2 The strong claim of incapacitation

Because separating incapacitation from deterrence is usually hard, only a few articles have tried to estimate the effect of prison on crime through incapacitation.

The article by Buonanno and Raphael (2013) is probably the strongest one showing an incapacitation effect on crime. They look at the 2006 Italian collective pardon that released a large number of prisoners unexpectedly. This shock increased crime (primarily theft) at a rate of about 20 crimes per released prisoner per year. (They net out the small deterrence effect from enhanced sentences for some released inmates that Drago et al. (2009) studied; see Section 23.3.2.2.) Barbarino and Mastrobuoni (2014) find similar size

⁴⁹ This is not necessarily true. If, for instance, incarceration rose because all sentences were doubled, the characteristics of the average prisoner would not have changed. But Section 23.5.1 shows that part of the reason for greater imprisonment is more admissions, particularly for less serious crimes.

effects from incapacitation when they study earlier collective pardons in Italy. Notice that we have no information on who committed these 20 additional crimes; we know only that they occurred. The released prisoners could have been victims; they could have committed more crimes than this but squeezed out other offenders; they could have committed fewer crimes than this but spread crime like a contagion.

Provinces with smaller initial inmate populations experienced larger increases in crime per released inmate, a finding consistent with diminishing marginal returns from incapacitation. [Johnson and Raphael \(2012\)](#) also find diminishing marginal returns from incapacitation between American states, but with a more complex identification strategy. [Liedka et al. \(2006\)](#) find the crime-prison elasticity falling to zero with an incarceration rate between 325 and 425 per 100,000. Diminishing marginal returns imply that crimes per prisoner released in Italy, which had an incarceration rate below 100 per 100,000 population after the collective clemency, are likely to be greater than crimes per prisoner in the early twenty-first century United States.

[Vollaard \(2013\)](#) also looked at a natural experiment in a jurisdiction with a low rate of imprisonment. In the early twenty-first century, the Netherlands introduced an enhanced sentencing policy for habitual offenders—individuals with more than 10 prior convictions. For these offenders, sentences for burglary and other property crime rose from about 2 months on average to about 2 years. Most of the individuals affected were older drug users; the average number of prior convictions for the affected group was over 30. Vollaard used variation across cities in the phase-in of the law—plausibly exogenous—to estimate large decreases in property crime (the affected offenders engaged in very little violent crime before the policy was implemented). He finds sharply diminishing returns. He does not, however, try to determine how much of the crime reduction is due to deterrence rather than incapacitation.

[Levitt \(1996\)](#) uses prison overcrowding litigation as an instrument for decreases in the state prison population. If this litigation does not affect potential offenders not in prison—for instance, because they do not know about it—but does affect prosecutors, judges, and parole boards, then the effect measured should be purely one of incapacitation. Levitt finds a reduction of 15 index crimes per year of imprisonment, and an elasticity of 0.4 for violent crime and 0.3 for property crime. The crimes per prison-year estimate is around that found in the inmate surveys. Considerable debate about Levitt's instruments has followed this article ([Durlauf and Nagin, 2011](#), p. 52). For instance, if prison overcrowding litigation reduces overcrowding, then prisons will be less unpleasant, and potential offenders may learn this from contacts in prison, whether they know about the litigation or not.

[Kuziemko and Levitt \(2004\)](#) estimate incapacitation effects for a specific group of prisoners, those sentenced for drug offenses. They show that imprisoning drug offenders has a sizable effect on violent and property crime—about as big as the effect of imprisoning violent and property criminals. Since imprisoning drug offenders should not deter violent

and property crime, and since increases in violent and property crime do not tautologically lead to greater imprisonment for drug crime, these estimates are plausibly incapacitation effects. But some questions remain. Perhaps greater drug imprisonment is caused by more police; then violent and property crime should be deterred by the same thing that is causing greater drug imprisonment. Moreover, Kuziemko and Levitt show that drug crime imprisonments lead to higher drug prices, and higher drug prices probably reduce demand. If the volume of drug use falls, the violent and property crime that surrounds illicit drug use may fall too. So the Kuziemko–Levitt estimates cannot be interpreted as pure incapacitation effects.

The coincidence between the [Levitt \(1996\)](#) estimate of the effect of prison-years on crime and inmate survey results for the weak claim makes it tempting to argue that the weak claim implies the strong claim: crime goes down by the number of crimes prisoners would have committed. [Bushway and Paternoster \(2009\)](#), for instance, cite the literature on hot-spot and terrorism-induced policing and say that because crimes prevented by policing are not displaced, crimes prevented by imprisonment are not likely to be displaced either.

This claim is based on some dubious reasoning. Suppose Mr A and Mr B are walking down Main Street on their way to an economics lecture, Mr A at 8:10 am and Mr B at 8:15 am. Both are moderately larcenous. Mr C has left his bicycle unlocked outside 100 Main Street. Everything else being equal, Mr A will steal the bicycle, and Mr B will not. If a police officer is standing at 102 Main Street, neither Mr A nor Mr B will steal the bicycle, and one crime will be averted, with no displacement. But if Mr A happens to be in prison that morning, and no police officer is stationed nearby, Mr B will steal the bicycle instead; the crime will be displaced. Similarly, a gang may choose to lay low if police occupy its territory, but they will replace a member who is in prison. Personal displacement and spatial displacement are different phenomena.

Of course, displacement may be negative as well as positive: an individual's release from prison may cause more crime than he/she commits himself/herself. The simplest example of this is the released prisoner who is shot down by old foes immediately on his/her return home. Models of preemptive murder such as that of [O'Flaherty and Sethi \(2010c\)](#) imply that small changes in the distribution of the propensity to commit murder can set off spirals of tension that result in large changes in actual murder rates.

Using evidence from the 2006 Italian collective pardon discussed above, [Drago and Galbiati \(2012\)](#) found that the average disincentive that an individual's peer group from prison faced had about the same effect on his recidivism that his own disincentive did. Release of one prisoner can affect the criminal behavior of many people.

23.5.2.3 Incapacitation plus deterrence effects

Starting with [Ehrlich \(1973\)](#), many studies have regressed crime on the prison population. [Donohue \(2009\)](#) surveys many recent studies, and concludes that they find an

elasticity of crime with respect to the state prison population of around 0.15–0.30. This elasticity translates into 15–30 crimes averted per prisoner-year in 1977, and four to seven crimes averted per prisoner-year in 2005 (Donohue, 2009, table 9.2).

Durlauf and Nagin (2011) maintain that because the independent variable in these regressions, prison population, is endogenous, the correlations have no obvious interpretation. This view is based on the retrospective story of incarceration that arises if prison is solely about deterrence and retribution: the prison population depends on past crimes, then, and the rules and policies under which past crimes are dealt with. The alternative, prospective view of incarceration (a view more in keeping with incapacitation and rehabilitation) is that authorities decide how many people they can afford to imprison and roughly who these people are, and then find pretexts to arrest and hold these people. The widespread use of prospective criteria in sentencing, probation, and parole decisions, as well as the sensitivity of the prison population to budget constraints (Spelman (2005), for instance, uses budget data as an instrument for incarceration), is evidence for the prospective rather than the retrospective theory of incarceration. To the extent that past crime is only a pretext for incarceration, the Durlauf and Nagin critique carries less force. More research into the positive determinants of the prison population seems in order.

23.5.3 Postprison effects

Prison may change people—their opportunities as well as their personalities—and these changes almost certainly affect their propensities to commit crimes. If the average prisoner serves 3 years (an overestimate) and then lives another 30 years, 90% of it free, then in the steady state the number of former prisoners is 4.5 times as great as the number of current prisoners. Postprison effects on crime are likely to be large relative to in-prison effects. (If incapacitation reduces crime by nine crimes per prisoner-year, but prison experience increases crime by two crimes per person-year when free, then in the steady state prison has no effect on crime. But increases in the prison population would cut crime and decreases would increase crime, and so regressions would say it was effective.)

Prison could either increase or decrease crime among former prisoners. Since prisoners are not randomly selected, former prisoners are not randomly selected either, and finding causal effects of prison is hard.

Prison experience could make people less likely to commit crimes in several ways. The first is what criminologists call “specific deterrence”: the experience of being caught, convicted, and sent to prison could make people update their priors about whether they will escape punishment for their crimes. As we noted above, Lochner (2007) and other articles in the specific deterrence literature find modest effects of this nature. Prison experience may also alter people’s views of what prison is like, but the direction of the effect is ambiguous: prison could be either less or more odious than anticipated.

Prison could also give people a chance to make traditional human capital investments, since the opportunity cost of their time is low. In the early 2000s, 31% of state prison inmates participated in vocational programs and 38% participated in educational programs (Lerman, 2009, p. 152). Higher legitimate wages after prison would reduce the incentive to commit crime. Prisoners may also make investments in drug and alcohol treatment, social and behavioral adaptation, and mental health care. They may become more fervent in their practice of religion, or may convert to a different faith.

On the other hand, the stigma of a prison record hurts employment prospects in legitimate jobs, because of both legal prohibitions and employer beliefs. Prisoners also lose employment contacts and skills while in prison, and may accumulate child support debt that produces extremely high marginal tax rates on reported earnings. The net effect of prison on employment and earnings is probably negative; see Holzer (2009) for a survey, and the many contributions in Bushway et al. (2007) for a variety of perspectives on this issue.

However, two recent studies that used random assignment of criminal cases to judges with different sentencing propensities as an instrument for imprisonment have found no effect on employment. Kling (2006) compares sentences of various lengths to see whether additional prison time harms subsequent earnings, and Loeffler (2013) examines the difference between no prison time (e.g., probation) and some positive amount of it. These are both local average treatment effects, but they are in different neighborhoods in the distribution of prison spells. They do not rule out an employment effect from arrest rather than imprisonment.

Among some populations, prison is viewed as a rite of passage. As Lisa Ling observes in a 2006 documentary on Mara Salvatrucha, a criminal gang with a reputation for extreme brutality: “In the surreal world of MS-13, prison isn’t punishment—It’s finishing school.” Such cases in which incarceration loses its deterrent effect completely are extreme and probably rare, but even for more routine offenders the propensity to commit crimes can increase as a result of having served time in prison. For instance, Bayer et al. (2009) show that a person who commits a certain type of crime (say, burglary) before prison is more likely to commit that crime after prison if he serves more time with others who have committed that crime.

Prison may also alter inmates’ personalities. “[A] harsh prison environment can make inmates less social, more violent, more likely to internalize the stigma associated with being a criminal” (Lerman, 2009, p. 153). The wikiHow “How to Deal with Being in Prison” advises: “It’s better to get into a fight and lose than to be seen as cowering or placating. . . if an altercation is unavoidable, react quickly and with aggression.”⁵⁰ Similarly, the wikiHow “How to Survive in Federal Prison” advises: “Don’t trust anyone. That goes for guards, prison officials, and the person in the cell next door. If someone is

⁵⁰ Accessed on September 20, 2013 at <http://www.wikihow.com/Deal-with-Being-in-Prison>.

being nice to you. . . They almost always have some hidden motive that you don't know about. In prison, nothing is free."⁵¹ Most people cannot alter their personalities quickly or at will. A personality developed for coping with prison life may easily lead to problems in civilian life. The traits that these documents advise prisoners to cultivate—aggression, preemptive violence, mistrust—can easily lead to violence and criminality in civilian life; they are almost precisely the traits that the social and behavioral interventions described in [Section 23.2.2.2](#) try to eliminate.

Empirical studies of postprison offending experience are of two kinds: some look at particular ways that incarceration can be structured, and others compare incarceration with some form of noncustodial supervision (or no supervision whatsoever). Rehabilitation works in the sense that some activities can reduce recidivism among those prisoners who participate relative to those who do not: vocational programs, cognitive behavioral therapy, and drug treatment, for instance ([Bushway and Paternoster, 2009](#)). Other activities increase recidivism: high-security imprisonment for prisoners with little prior criminal involvement, for instance ([Chen and Shapiro, 2007](#); [Lerman, 2009](#)). These studies cannot answer the question of whether people who go to prison and participate in effective programs are more likely to commit future crimes than otherwise identical people who do not go to prison.

The second question, whether the average prison experience is criminogenic relative to some form of noncustodial supervision, has been the subject of hundreds of studies, mostly outside economics, and several meta-analyses. The results of these studies are mixed and not definitive ([Lerman, 2009](#), pp. 153–154), although a review by [Nagin et al. \(2009\)](#) concludes that the effect of prison is probably weakly criminogenic, or null ([Loeffler, 2013](#)).

Two recent studies in economics with careful identification strategies have added to this literature. [Aizer and Doyle \(2013\)](#) find that juvenile detention greatly reduces the probability of high school graduation and increases recidivism. [Di Tella and Schargrodsky \(2009\)](#) show that assignment to incarceration rather than electronic monitoring raises recidivism moderately among adults in Argentina. Both studies use random assignment to more or less lenient judges as an instrument, and so implicitly study defendants close to the margin of imprisonment. The range over which these results generalize might therefore be limited.

The mixed results should not be surprising. Prisoners differ, prisons differ, and non-custodial experiences differ. So the effect of prison on future offending should not be a constant.

The final type of postprison effect is not on prisoners, but on their children. [Johnson \(2009\)](#) calculates that roughly one-fifth of black children have a father who has been incarcerated. Young children whose parents are or were incarcerated exhibit more

⁵¹ Accessed on September 20, 2013 at <http://www.wikihow.com/Survive-in-Federal-Prison>.

behavioral problems than other children, and Johnson shows that this effect persists after adding many more controls—although he cannot establish causality. Behavioral problems as a child are correlated with criminal offending as a young adult. [Wildeman \(2010\)](#) also finds that paternal incarceration is associated with more physically aggressive behavior by boys. Like Johnson, he cannot establish causality, but uses several methods (propensity scores, fixed effects, and placebo regression) that strongly suggest it.

23.5.4 Prison gangs

Prisons are dangerous and inhospitable places. Guards cannot enforce peaceful coexistence among prisoners. Prisons contain a higher proportion of psychotic and violent individuals than normal communities do, and offer fewer opportunities for self-protection. In prison, one cannot choose whom to dine, exercise, sleep, or shower with. One cannot carry a mace or a gun, or move to a safer environment. State-sponsored deterrence is also less effective in prisons: prisoners cannot be threatened with immediate loss of freedom or access to possessions or to friends and sexual partners; or with the stigma of being known as convicts; or with the risk of losing a good job. Prisoners cannot buy many goods that they want, and guards will not enforce most agreements among prisoners, especially those dealing with trade in contraband.

Each prisoner's experience then depends on how well he/she interacts with other prisoners. When prisons were small and uncrowded, they were governed by convict norms that reduced violence and allowed trade in contraband to flourish. These norms broke down as prisons expanded, transfers became more frequent because prison systems were larger, prisoners became younger, the population became more heterogeneous, and as the influx of drug users and drug dealers made the trade in contraband more lucrative. To enforce order, prison gangs replaced prison norms ([Skarbek, 2012](#)). Gangs strong enough to protect their members were also strong enough to act as predators toward nonmembers.

Prison gangs may affect crime outside prison. When prisoners leave prison, they do not necessarily leave the gang. Gang membership provides them with trusted associates and contacts. Since gangs import illicit drugs into prison, these contacts include drug-trafficking organizations. Thus gang membership gives former prisoners a comparative advantage in drug dealing and other types of crime. For instance, the Mexican Mafia (La Eme), originally founded in prison, became a drug-trafficking organization outside prison ([Fleisher and Decker, 2001](#)). The finding of [Drago and Galbiati \(2012\)](#) on peer effects among Italian ex-prisoners is consistent with this interpretation.

Whether encouraging street gang membership increases or decreases crime, especially index crime, is unclear. On one hand, the requirements of gang membership may force individuals who have joined or who want to join a gang only for insurance purposes to commit more index crimes than they otherwise would. On the other hand, gangs,

especially large gangs, may create monopolies in illicit drug markets and so reduce the number of disputes that need to be settled violently; La Eme, when it was originally formed, drew on the membership of a large number of street gangs (Skarbek, 2012). How the prison boom altered the organization of criminal activity outside prison is a topic that has not been well studied.

23.5.4.1 Preprison effects and prison culture

The influence of prison gangs is not necessarily confined to ex-prisoners; it may affect “preprisoners” too. Just as La Eme started as a prison gang and became a street gang, the Crips started as a street gang and became a prison gang. Entering prison without friends or gang connections is probably dangerous and definitely scary. A person who has a high probability of going to prison sometime in his/her life may be well advised to establish gang ties before that occurs.

Anticipation of prison may also cause people to cultivate personality traits that are dysfunctional outside prison, such as the aggression and suspicion that we described in Section 23.5.3. Another prison trait relevant to crime in the outside world is unwillingness to cooperate with authorities: “Do not snitch. If you see something illegal or violent, walk away and do not divulge any information if questioned later. If you become known as a snitch, other inmates will target you as retaliation.”⁵²

A community outside prison filled with people who have internalized these ways of interacting is a dangerous and crime-ridden place, and one where those who do not act like “prisoners” may find it in their best interest to start acting that way, even if they are neither former prisoners nor preprisoners. Prison culture—not just prison-inspired fashions—can spread. Standard econometric techniques would not detect a rise in crime caused by a spread of prison culture. Spikes in incarceration would still cut crime, and individuals with actual prison experience would not look or behave much differently from individuals without actual experience.

Cooperation with law enforcement is risky and costly in many communities. Furthermore, like the willingness to kill discussed in Section 23.4.5, the willingness to provide evidence or testify in court is characterized by strategic complementarity. Incarceration substantially reduces the ability of a defendant to harm witnesses, and a conviction becomes more likely if multiple witnesses step forward. Since the costs of cooperation to an individual depend on the incidence of cooperation in the community, multiple equilibria can arise: high levels of cooperation may be sustainable, but there may also exist equilibria with *collective silence*: no witness comes forward because no witness expects others to come forward to corroborate testimony (O’Flaherty and Sethi, 2010b).⁵³ Under

⁵² How to Deal with Being in Prison (wikiHow): <http://www.wikihow.com/Deal-with-Being-in-Prison>. Accessed on September 20, 2013.

⁵³ For a spectacular example of collective silence, see Kocieniewski (2007).

these conditions the problem of identifying the effects of mass incarceration on crime empirically becomes especially acute.

Because prisoners come from, leave, and return to neighborhoods that are spatially concentrated, incarceration is thus likely to have spatial effects. Children growing up may anticipate a different life course and so invest differently, adolescents may encounter young adults who are more aggressive and mistrustful, police may find citizens less cooperative, women may find fewer prospective partners and so lose bargaining power in any relationships that do form, children in the next generation may find their lives disrupted by paternal incarceration and grow more aggressive, and their peers will have to deal with their aggressiveness. In economics, a full spatial model has not been investigated yet, either theoretically or empirically, but criminologists such as [Clear \(2007\)](#) are investigating these consequences. Such neighborhood effects might be the reason why some recent studies have found little or no effect of individual imprisonment on earnings or criminality. [Loeffler \(2013\)](#) writes: “If many prisons are simply extensions of high-disadvantage neighborhoods, then the effects of moving between these two environments might be smaller than previously thought.”

23.5.5 Trade-offs

Although the consensus of studies is that the average effect of incarceration is to reduce crime, and the marginal effect may also be to do so, the current volume of incarceration is probably not an efficient way of reducing crime.

[Cook and Ludwig \(2010\)](#) make this case simply. Suppose early twenty-first century sentencing and parole practices were scaled back to 1984 levels. Following [Raphael and Stoll \(2009b\)](#), they say that this would reduce the state prison population by about 35% of the increase since 1984, and state prison expenditures would fall accordingly.⁵⁴ If the federal prison population also fell by 35% of the increase since 1984 and expenditures fell accordingly, the government budget savings would amount to \$12 billion a year, but the smaller prison population would raise violent index crimes by 26,000 a year and

⁵⁴ The thought experiment that Cook and Ludwig ask for is not the thought experiment that went into calculating the data in [Table 23.1](#). For that table the crime rate and arrests per crime were held at 1984 levels, and the question that was asked was what would happen to steady-state imprisonment if the time served went to 2002 levels. Cook and Ludwig hold the crime rate and arrests per crime at 2002 levels and reduce the time served to 1984 levels. The calculation that would arrive at 35% uses all categories of prisoner, including parole returnees. For index crimes, steady-state imprisonment per capita was 201.8 in 1984 and rose to 305.3 in 2002, an increase of 103.5. Holding the crime rate and arrests per crime at 2002 levels and reducing time served to 1984 levels produces an imprisonment rate of 245.0. So the proportion of the increase that would be eliminated is 58.3%. For drug offenders and index crime offenders together, the same sort of calculation implies that going to 1984 time served would reduce the increase by 44.7%. Federal prisons have a higher proportion of drug offenders than state prisons do, and a larger proportion of their increase is due to drug offenders, so the proportion of the federal prison increase due to greater time served is probably considerably less than that for state prisons.

property index crimes by 186,000 a year in the absence of any compensating policy change. But shifting that \$12 billion in government spending to police, preschool for poor youngsters, or social-behavioral skills programs for troubled adolescents would decrease index crime by substantially more than these amounts.

This last claim confuses results on the weak claim of criminogenic characteristics—which has been established for preschool and social-behavioral skills—with the strong claim of criminogenic characteristics—which has not been established for these interventions. Only the strong claim is relevant for this calculation. Still, the comparison for police is sufficient to make the point that prison expansion is not the best use of money, and other interventions for which the strong claim has been established such as psychopharmaceuticals could be substituted for preschool and social-behavioral training.

These cost calculations ignore any changes in crime from altered recidivism, preprison incentives, or spread of prison culture. They also ignore the benefits that prisoners themselves would receive from not being incarcerated, as well as the benefits to those who are “knitted together with offenders in networks of social and psychic affiliation” (Loury, 2009). The costs to prisoners and their families are probably large, but are often ignored, or calculated as loss of wages. Donohue (2009), in his magisterial cost-benefit analysis, devotes 10 pages to the elasticity of crime with respect to imprisonment and another 14 pages to the cost of crime, and cites many studies in these pages; he devotes a paragraph with no citations and part of a footnote to the direct current cost of incarceration to prisoners.

The willingness of those with the means to pay large fines and incur large legal expenses to avoid prison suggests that for most people going to prison or jail is much, much worse than losing earnings. Cook (1983) argues for counting the costs to prisoners, but the practice is not widespread. We know of no study that attempts to estimate the cost of prison to prisoners or their families.

When sentencing and parole decisions are based on retribution or deterrence, arguments can be made for ignoring costs to prisoners. But ignoring prisoner costs is much harder to defend when these decisions are based on considerations of incapacitation—that is, crimes not yet committed. Incapacitation, after all, could be achieved just as well by sending likely offenders to Club Med as it could be by sending them to prison.

23.5.6 Race and mass incarceration

Many other government policies—agricultural subsidies, for instance—are monumentally inefficient, but few have excited the passion among distinguished intellectuals that mass incarceration has. Economists should understand this passion.

Prisons are profoundly racial places. Even though a substantial minority of prisoners are non-Hispanic whites, the prison population in the United States is dominated by minorities. In fact, prisons *make* people black: some individuals who are white outside

prison are black inside prisons, both as reported by themselves and as reported by others (Saperstein and Panner, 2010). Prisons are also highly segregated; prison gangs, for instance, are typically organized along racial and ethnic lines. The wikiHow on federal prisons recommends: “It is crucial to your survival to immediately show your allegiance to your race. . . If you are some white suburbanite 19-year-old that pledged yourself as a crip, and you used to buy the dope you got busted selling from your homeboys in the projects, that doesn’t mean you can link up with them in prison. If you’re white and you walk in slapping high fives with the brothers before you shake hands with the white dudes, you’re going to send a rift through the whole community.”⁵⁵

Physical coercion, dishonor, and loss of freedom—the hallmarks of prison—are central to the long and tumultuous history of race in the United States. Alexander (2010, p. 2) writes: “Today it is perfectly legal to discriminate against criminals in nearly all the ways it was once legal to discriminate against African Americans. Once you’re labeled a felon, the old forms of discrimination—employment discrimination, housing discrimination, denial of the right to vote, denial of educational opportunity, denial of food stamps and other public benefits, and exclusion from jury service—are suddenly legal. As a criminal, you have scarcely more rights, and arguably less respect, than a black man living in Alabama at the height of Jim Crow. We have not ended racial caste in America; we have merely redesigned it.”

Loury (2008) explores the prevailing political reaction to the prison boom—the question of what really determines the level of incarceration. The widespread complacency in the face of stark racial disparities in incarceration rates, according to Loury, has deep historical roots. Following Patterson (1982), he argues that American slavery was “not merely a legal convention but also a superstructure of justifying ideas defining and legitimizing an order of racial hierarchy,” and that the profound “racial dishonor” that was part of this superstructure of ideas could not be eliminated simply by a change in the legal status of slavery. Hence, there survived “an entrenched if inchoate presumption of inferiority, of moral inadequacy, of threat to public safety, of unfitness for intimacy, of intellectual incapacity” when members of the stigmatized group were viewed by those outside it. Without this contempt, argues Loury, the staggering racial disparities in mass incarceration would not be considered socially acceptable, and would therefore not be politically sustainable.

This perspective also sheds light on the deep ambivalence toward aggressive enforcement that is evident in the inner-city communities where victimization is highest. These victims do not share the contempt for young black males or the “enthusiasm for their debasement” that is felt by those who reside at some distance, both social and geographic, from these communities. Furthermore, it helps explain why significant racial disparities in

⁵⁵ How to Survive in Federal Prison (wikihow): <http://www.wikihow.com/Survive-in-Federal-Prison>. Accessed on September 20, 2013.

drug arrests are tolerated, even while disparities in drug use are negligible. And it sheds light on the significant disparities in the sentencing of convicted killers, with defendants considerably more likely to receive a death sentence when the victim is white. Loury argues that vigorous enforcement policies have sometimes been politically popular precisely *because* they have a disproportionate impact on blacks: “Institutional arrangements for dealing with criminal offenders in the United States have evolved to serve expressive as well as instrumental ends” (Loury, 2009).

As we have discussed at length in this chapter, there are many instances in which membership of a visibly identifiable group affects the incentives faced by offenders and victims engaged in a criminal interaction. Black offenders may face less resistance, for instance, if they are perceived to be more desperate. This makes robbery more lucrative, and affects the relative rates of offending across different identity groups (see [Section 23.4.3](#)). The disparity in this case is driven by different incentives, rather than differences in nature or deep cultural characteristics. But these incentives remain hidden from most observers, which can result in what Loury calls “essentialist causal misattributions.” That is, the disparities are attributed to the prevalence of largely immutable criminogenic characteristics in a population, rather than the recognition that they result from the structure of incentives and interactions. As a result, the disparity does not result in urgent calls for corrective action.⁵⁶

23.6. BIG SWINGS IN CRIME

Several times in recent memory—that is, in the era of reasonably reliable data—the rates of almost all kinds of crime have moved up or down together over large geographic areas for a decade or so, as if they were all moved by a single gigantic hand. The best known and best studied such incident was the great American (and Canadian) crime decline of the 1990s. Other incidents are the great American crime rise from the mid 1960s to the mid 1970s, and the crime declines in the United States and most other rich countries in the early 2000s.

These swings are large; changes of more than 50 log points are common. They are thus typically much larger (though less frequent and timed considerably differently) than the swings in output or employment that macroeconomists concentrate on. They are no less worthy of study.

⁵⁶ Loury (2002) illustrates this point by comparing gender disparities in two domains: science classes and prisons. Males are heavily over-represented in both domains but only in the former is there a sense that corrective action is required, because “it offends our basic intuition about the propriety of underlying social processes that boys and girls have different levels of achievement in the technical curriculum of our schools.” The fact that racial disparities in incarceration rates does not offend our intuitions in the same manner is attributed to the legacy of racial dishonor.

The synoptic literature on these events is limited to the American crime decline of the 1990s: articles by [Levitt \(2004\)](#) and [Blumstein and Wallman \(2006b\)](#), and books by [Blumstein and Wallman \(2000, 2006a\)](#) and [Zimring \(2007\)](#). Levitt and Zimring do, however, discuss the great American crime rise. Many other articles attempt to explain pieces of the American 1990s; at least one of them, by [Reyes \(2007\)](#), says something about the crime rise and the twenty-first century.

In this section, we will first establish some stylized facts about the four big swings, and then review the explanations for the American crime decline of the 1990s. We will then look at the other three big swings, and try to extend those explanations. We begin with the 1990s because this episode has received the most study. Then we look at the 2000s because they appear to be related to the 1990s.

Our conclusion is that we do not know now what caused most of these big swings, but we should. This does not mean that the factors traditionally studied—police, prisons, the macroeconomy, and demographics—are important or that studying them is useless. Something (probably several things) is causing these big swings, we do not know what it is yet, and it is more powerful than the traditional factors.

23.6.1 Stylized facts

[Figure 23.1](#) shows US rates for reported offenses for four crimes from 1960 to 2012: murder, robbery, burglary, and motor vehicle theft. We set 1986, the midpoint, equal to 100

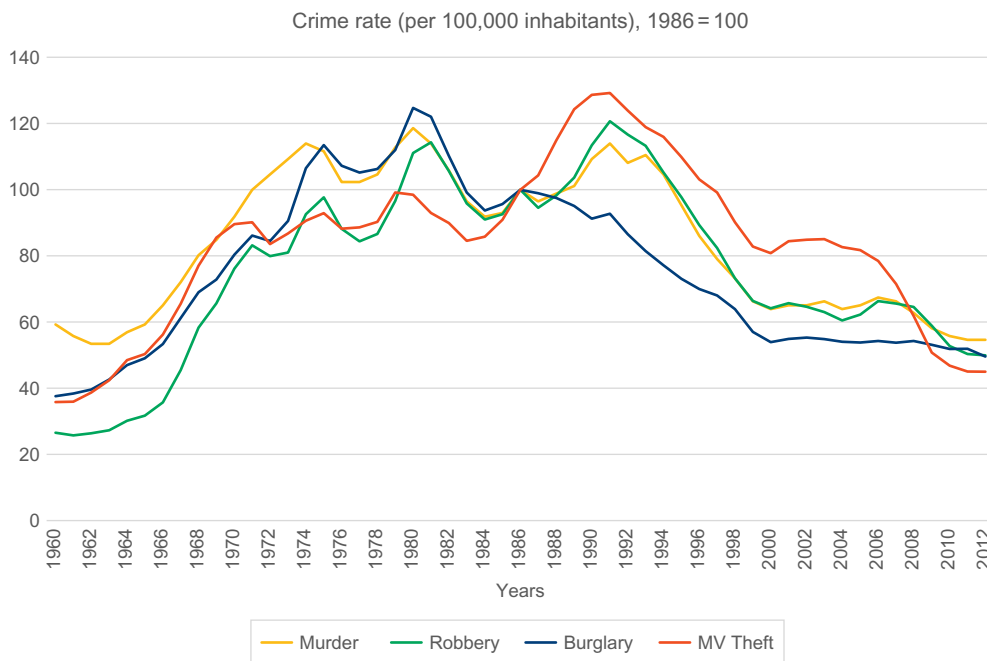


Figure 23.1 Historical swings in crime rates.

for each crime. Murder and motor vehicle theft are quite well reported in the United States. Rape, aggravated assault, and larceny-theft, the index crimes we have omitted from this figure, suffer from particularly large reporting problems, especially over a long period of time.

The great rise and the great decline are both easily evident in [Figure 23.1](#), as is the continued decline in the twenty-first century. The great rise and the great fall are of roughly the same magnitude, although, of course, the great rise is much bigger relative to its starting point. The twenty-first century decline is not as dramatic as its predecessors (except for motor vehicle theft), but is still a noteworthy event.

[Figure 23.1](#) deals with reported crimes, but when available, victimization statistics follow the same pattern. The National Crime Victimization Survey (NCVS) did not begin until the end of the great crime rise, and so cannot be used as a comparison for that event; and the NCVS never gives data on murder (for obvious reasons). But otherwise, the NCVS and the Uniform Crime Reports (UCR) both show across-the-board declines in the 1990s and 2000s.

What about other developed countries? [Zimring \(2007\)](#) compared the US experience in the 1990s with that in other G-7 countries (he omitted Germany because reunification was ongoing in 1990). Canada mirrored the United States, but the other countries he looked at—France, Italy, Japan, and the United Kingdom—did not. He concluded that there was no typical pattern, and [Levitt \(2004\)](#) reaches a similar conclusion.

But the twenty-first century has a trend, and it is downward. This is evident in [Table 23.2](#), where we look at a wider set of countries—OECD members with 2010 gross national income per capita above \$25,000 (purchasing power parity)—for the four crimes in the early twenty-first century.⁵⁷ Among the G-7 countries in this period, the rates of almost every crime in every country went down, and sometimes by spectacular margins. The only exception in the G-7 may be the robbery rate in Italy, but reporting issues make even this exception unclear.⁵⁸ Over the longer period, the United States appears not as an outlier, but as a precursor.

⁵⁷ In [Table 23.2](#) the period for murder is 2000–2011. The period for all other crimes is 2003–2011. Exceptions are as follows: For Austria, Spain, France, Greece, and the Netherlands, 2004 motor vehicle theft is substituted for 2003 motor vehicle theft. For Italy, New Zealand, Sweden, and the Netherlands, 2005 motor vehicle theft is substituted for 2003 motor vehicle theft. For Canada, 2010 motor vehicle theft is substituted for 2011 motor vehicle theft. For Austria, Japan, and Israel, 2004 robbery is substituted for 2003 robbery. For Spain, 2005 robbery is substituted for 2003 robbery. For Austria and Israel, 2004 burglary is substituted for 2003 burglary. The UK data are for England and Wales; the population for crimes other than murder is from the UK Office for National Statistics, Statistical Bulletin, Population Estimates for England and Wales Mid-2002 to Mid-2010 Revised (National), 13 December 2012.

⁵⁸ Italy introduced a new system of crime reporting in 2004 and progressively implemented it. The [European Sourcebook of Crime and Criminal Justice Statistics \(2010, p. 140\)](#) states: “This change implies a risk in comparing police statistics among 2004, 2005, and 2006.” In the United Nations Office on Drugs and Crime statistics, robbery rises tremendously in 2004, and the burglary category disappears.

Table 23.2 Rates of change for four reported crimes per capita in the 2000s

	Murder	Robbery	Burglary	Motor vehicle theft
Australia	−31.3%	NA	−45.9%	−50.6%
Austria	−20.0%	−17.6%	−40.3%	−49.8%
Belgium	−14.3%	−14.4%	1.2%	−54.2%
Canada	−6.3%	−14.7%	−41.6%	−56.5%
Denmark	−27.3%	−0.7%	−11.1%	NA
Finland	−4.3%	−23.2%	−33.1%	NA
France	−33.3%	−8.0%	−20.0%	−3.7%
Germany	−33.3%	−19.4%	−6.7%	−28.2%
Greece	128.6%	178.9%	172.6%	103.2%
Iceland	−50.0%	−4.4%	−41.3%	NA
Ireland	−10.0%	51.3%	−2.8%	−25.4%
Israel	−16.7%	−1.4%	−53.0%	NA
Italy	−30.8%	32.9%	NA	−14.6%
Japan	−40.0%	−50.0%	−62.2%	−50.1%
Korea	30.0%	−46.8%	NA	NA
Luxembourg	−11.1%	NA	NA	NA
Netherlands	NA	−28.0%	NA	−24.8%
New Zealand	−30.8%	NA	NA	−19.0%
Norway	109.1%	5.4%	−45.9%	−49.6%
Slovenia	−55.6%	13.1%	−22.7%	−30.9%
Spain	−42.9%	−22.8%	NA	−53.7%
Sweden	−10.0%	7.2%	−27.3%	−51.5%
Switzerland	−40.0%	14.5%	−9.8%	−89.3%
United Kingdom	−37.5%	−32.2%	−42.5%	−70.3%
United States	−14.5%	−20.4%	−5.5%	−47.2%
Observations	24	22	19	19
Number increasing	3	7	2	1
Median	−23.6%	−11.2%	−27.3%	−49.6%

NA, not available.

Source: United Nations Office on Drugs and Crime.

23.6.2 Explaining the American crime drop of the 1990s

It is convenient to think of two classes of explanatory variable for any change in crime rates: traditional and nontraditional. The traditional category is composed of criminal justice variables (police and prisons), demography, and macroeconomics. “Nontraditional” refers to everything else. The consensus is that police and prisons by themselves are insufficient to explain the crime drop of the 1990s. Most articles conclude that all the traditional variables together are also insufficient, although a few articles maintain that this combination can explain a large proportion of the drop in property crime in the 1990s. Articles that claim to explain all or most of the crime drop of the 1990s—these are generally in economics—appeal to a variety of nontraditional

explanations, while synoptic works in criminology are usually content to leave a great deal of the crime drop unexplained.

In this section, we will start with the traditional variables, and add up plausible estimates of the size of their impacts. Then we will look at a variety of nontraditional variables. The latter discussion will be considerably more speculative than the former.

23.6.2.1 Traditional explanations

Police force size

Levitt (2004) states that the number of police per capita rose 14% over the period 1991–2001. As Chalfin and McCrary (2013) point out, police are hard to count on a local basis, and national estimates yield new problems with the classification of agencies and broad geographic reach. But there is little question that US police force size grew substantially during that decade.

How much police force size mattered is controversial. Criminologists generally cite earlier studies and argue for little or no effect; Eck and Maguire (2000) provide a good survey of this literature. Economists work with later articles with good identification strategies. Levitt uses his 2002 article, from which he interprets an elasticity of -0.4 for all types of crime with respect to the number of police.

This estimate seems outdated. Chalfin and McCrary (2013) provide a helpful review of estimates from several more recent articles. For murder, it seems clear that an elasticity over -0.4 is appropriate; estimates for post-2002 articles range from -0.50 to -2.73 , with a median of -0.87 . For violent crime, the elasticities range from -0.34 to -1.13 , with a median of -0.66 (elasticities for robbery are much higher than those for rape or assault). Property crime elasticities, on the other hand, are probably somewhat smaller than the elasticity Levitt used (-0.4); the values in the literature range from $+0.11$ to -2.18 , with a median of -0.26 . The median estimates imply a decrease in murder of 11 log points, in violent crime of 9 log points, and in property crime of 3 log points.

Prison

The prison population rose 36% during the period examined by Levitt, and he uses an elasticity of -0.3 for murder and violent crime and -0.2 for property crime to attribute a 12% decrease in murder and violent crime to this cause (in log points, the contribution is 9), and a 7% decrease in property crime (6 log points). The elasticity is high relative to estimates in the literature; Donohue (2009), for instance, concludes from his survey that a figure half as high would be more reasonable; this would imply log point decline for murder and violent crime of 4 or 5.

These estimates are primarily a current incapacitation effect; as we noted in Section 23.5, they ignore the possibility of heightened offending by former prisoners. By the 1990s, the prison boom had been going on for long enough that the stock of former prisoners was high, and it grew during the decade.

Demographics

The consensus is that changing age composition contributed modestly to the crime decline of the 1990s. The proportion of the resident population aged 15–24 years fell from 14.9% in 1990 to 13.9% in 2000; the proportion aged 15–29 years fell from 23.5% to 20.8% (Zimring, 2007, p. 61). On the basis of a Blinder–Oaxaca decomposition of changes in the distribution of the population by 5-year age groups and correction for racial changes, Levitt (2004) estimated that changing demography decreased property crime by 5 log points and violent crime by 2 log points. He estimated a zero effect on murder.

Zimring (2007) suggests that demographics may have had a bigger effect than this, but he puts no number on his conjecture. His argument is based on a process of elimination: Canada experienced a crime decline similar to that in the United States, but did not add police or prisoners. That Canada had a demographic experience similar to that of the United States suggested to Zimring that demography might be more powerful than other analysts thought.

The idea of a nonlinear demographic effect—essentially, that big cohorts have higher age-specific crime rates—was first posited by Easterlin (1973). Steffensmeier et al. (1987), O'Brien (1989), and Levitt (1999) all argue that historical evidence does not support the existence of a cohort-size effect.

Prosperity

As far as macroeconomic performance is concerned, the 1990s is remembered as a good decade: unemployment fell and wages rose. Prosperity decreases crime (notice that the effect of prosperity on crime is a simpler question than the one we tackled in Section 23.3.3, concerning the effects of unemployment and wages on crime), except murder, and macroeconomic events almost certainly contributed to the crime decline of the 1990s. There is no consensus, however, about how much.

Levitt (2004) concentrates on unemployment and summarizes his understanding of the extant literature as implying that a 1% change in the general unemployment rate leads to a 1 log point change in property crime (in the same direction), and no change in violent crime. This implies a 2 log point decrease in property crime from prosperity in the 1990s (the general unemployment rate fell from 6.8% in 1991 to 4.8% in 2001).

By contrast, Freeman (2001) reviews the literature on prosperity and property crime, and explains how some of these articles differ in two ways from Levitt's approach. First, since most index crime is the province of low-skilled men, the relevant unemployment rate is the rate for that group, not the general rate. Second, wages matter as well as unemployment rates, and the wages that matter are those for low-skilled men.

The articles that Freeman uses do not consistently find an effect on violent crime or murder, but the implied estimates for property crime are considerably larger than the one Levitt used. Gould et al. (2002) probably give the largest impact estimate (they do not

Table 23.3 Estimated log point changes in crime due to traditional variables, 1991–2001

	Murder			Violent Crime			Property crime		
	Levitt	Smaller	Bigger	Levitt	Smaller	Bigger	Levitt	Smaller	Bigger
Police size	–5		–11	–5		–9	–5	–3	
Prison	–12	–6		–12	–6		–8	–4	
Demography	0			–2		–3	–5		–7
Macroeconomy	0		0	0			–2		–8
Total	–17	–11	–23	–19	–13	–24	–20	–14	–28
Actual (UCR)	–56	–56	–56	–42	–42	–42	–34	–34	–34
Residual	–39	–45	–33	–23	–29	–18	–14	–20	–6

have prison controls and do not compute instrumental variables estimates with police variables as controls; hence, we use their ordinary least squares estimate with police controls). From their article, we calculate an 8 log point decrease in property crime due to lower unemployment and higher wages for low-skilled workers.

Residual

The traditional variables do not explain all or even most of the crime decline of the 1990s. [Table 23.3](#) summarizes our analysis so far. For each type of crime, we start with Levitt’s estimated impacts, and then have columns for bigger or smaller estimates based on the discussion so far in this section. Unlike Levitt, we use log points rather than percentages to be consistent with the equations in the original articles.

Thus only for property crime, and only for the high estimates, do the traditional variables come close to explaining the crime decline of the 1990s. (This is somewhat consistent with [Imrohoroglu et al. \(2004\)](#), who claimed that they could explain the property crime decline over a longer period with enforcement and wage variables.) Otherwise, traditional factors always explain less than 60% of the crime decline, and maybe less than one-third. The unexplained residual is over 30 log points for murder, over 20 log points for violent crime, and around 10 log points for property crime.

Economists are used to seeing unexplained residuals, and are often quicker to label them than to explain them—thus, for instance, “technological progress” in growth regressions, and “discrimination” in Blinder–Oaxaca decompositions for racial or gender differences. In the case of crime, however, no handy label has appeared for the unexplained residual, and instead attention has turned to nontraditional explanations.

23.6.2.2 Nontraditional explanations

In this subsection, we will review many nontraditional explanations. Some of them appear to have merit, while others do not, and still others remain to be tested.

Capital punishment

Executions increased during the 1990s, although the number remained small relative to aggregate incarceration. [Levitt \(2004\)](#) takes the highest estimate in the literature of murders averted per execution and multiplies it by the decadal increase in executions; the result is a 1.5 log point decrease in murder. There is much reason to be skeptical of even this small impact, as no consensus exists on the effect of capital punishment on murder.

Abortion

[Levitt \(2004\)](#) attributes a 10% (11 log point) decline in crime across the board to the legalization of abortion (on top of demographic changes). As we concluded in [Section 23.2.1.5](#), this attribution is probably misplaced. There is an interesting correlation between events and dispositions in the early 1970s and crime in the 1990s, but that correlation does not appear to be acting through abortion. Part of the crime decline of the 1990s is due to this “Donohue–Levitt factor,” but we do not know yet what this factor is.

Crack

Crack appeared in US cities around 1985, and its introduction set off a round of violence, especially murders among young black men, that substantially raised crime rates ([Fryer et al., 2013](#)). The violence subsided in the early 1990s. Two stories have been told about why the violence subsided: either crack markets were cartelized and wars became less likely to break out among competing sellers, or demand fell when young generations saw the harm that crack had wrought on their older brothers and sisters.

[Levitt \(2004\)](#) and [Blumstein and Wallman \(2006b\)](#) attribute a large portion of the decrease in violent crime, especially murder, to the evolution of crack. In an accounting sense, this attribution is correct. Had the crack epidemic not occurred, or had it started in 1995 rather than 1985, the crime decline of the 1990s in the United States would have been smaller.

This accounting explanation, however, fails to address the fundamental question of why the crack epidemic occurred when it did, and why no similar epidemic has occurred since, either in the United States or in another developed country. Crack was a technological innovation in recreational drugs; surely it did not exhaust all possible technology in this field. Many innovations, such as methamphetamines, have spread in the world of recreational drugs since crack, and even more could have spread if the payoffs were large enough, but none have had anywhere near the impact on violent crime that crack had.

The United States experienced two recreational drug epidemics that apparently had large effects on criminal offending—crack in the late 1980s and heroin in the early 1970s. Both were centered in African American city neighborhoods, and both occurred when most varieties of index crime were close to their peak, as far as recorded history is concerned. Was this just an unfortunate coincidence, and was the fact that no such shocks occurred in the 40 years preceding 1970 or in the 20 years following 1990 just good luck?

For a day-to-day weather forecaster, hurricanes are random shocks too. But they do not often occur in the winter, and they do not often strike Nebraska. Meteorology explains these regularities about hurricanes. That crime-inducing recreational drug epidemics were concentrated in African American neighborhoods of US cities during a particular 20-year period is likewise a regularity that cannot simply be attributed to chance. Crack is something that needs to be explained; it is not an explanation.

Police productivity

Although improved high-profile police tactics are probably the most popular explanation for the crime decline, both criminologists and economists agree that they probably played at best a minor role. [Levitt \(2004\)](#) assumes that they are responsible for a 1 log point across-the-board decrease in crime, but he assigns this value without much conviction. Criminologists, except possibly [Zimring \(2011\)](#) on New York City, would be less generous.

The main reason for this skepticism is that there is no evidence of more effective police tactics being widely adopted. [Section 23.3.2](#) shows that hot-spots policing and terrorism-induced police presence reduce crime. There is no evidence that hot-spots policing became more widespread in the 1990s. The tactics the terrorism studies focus on—for instance, standing around in front of synagogues—are hardly innovative, and they probably did not proliferate more in the 1990s. During this period, many cities that adopted wildly different high-profile policing tactics simultaneously experienced large crime decreases.

Two innovations that have received considerable public attention have also failed to convince experts of their contribution to the crime decline—“broken windows” policing and Compstat. Broken windows policing grew out of a famous *Atlantic* article by [Wilson and Kelling \(1982\)](#) that hypothesized that general disorder in a neighborhood—broken windows that are not repaired—signaled to criminals that they could take over because nobody cared. Police could therefore reduce crime by combating disorder.

There are two reasons to believe that broken windows policing did not contribute significantly to the crime decline. First, the consensus is that it is ineffective: “There is widespread perception among policymakers and the public that enforcement strategies (primarily arrests applied broadly against offenders committing minor offenses) lead to reductions in serious crime. Research does not provide strong support for this proposition” ([Skogan and Frydl, 2004](#), p. 229). Second, broken windows was probably little applied. New York City is supposed to be the poster child for broken windows policing, but [Zimring \(2011\)](#) shows that it did not actually adopt this strategy: the vice squad was shut down, arrests for prostitution—one of the main examples [Wilson and Kelling \(1982\)](#) use—went down, and arrests for gambling went down after a brief rise

in the early 1990s. Marijuana arrests rose, but only among men. The New York City Police Department made pretextual arrests, not order-enhancing arrests.

By contrast, the problem with Compstat is that it has never been evaluated. Compstat is a management practice where top-level officials use real-time, small-area crime data to hold precinct commanders responsible for crime in their areas. It is a combination of hot spots and slack reduction. [Zimring \(2011, pp. 143–144\)](#) argues that Compstat was probably successful in New York City because the New York crime drop was unusually large, and there is no independent evidence elsewhere that Compstat does not work. He acknowledges that this argument is weak.

Hence, it is likely that no high-profile policing innovation was responsible for a significant portion of the crime drop. But could police have improved their performance in the 1990s by many small, widespread innovations—better communications, for instance, better use of data, more educated officers, more ambitious superiors, and so on? Productivity increases are accepted in the private sector without demonstration of the effectiveness of particular innovations.

If productivity increased in this fashion, then researchers who estimated the effect of police on crime would find a larger elasticity with data from the 1990s than with earlier data. No clean test like that is available. However, [Evans and Owens \(2007\)](#) used data from only 1990 to 2001, and so when [Chalfin and McCrary \(2013\)](#) compared their results with those of Evans and Owens, they reran their specification with data restricted to that period. Their main results are for the period 1960–2010. For violent crime, they estimated an elasticity of -0.34 for the full period, and -0.63 for the 1990s. For property crime, their estimates were -0.17 for the full period, and -0.31 for the 1990s. The police in the 1990s were more effective on average, as would be the case if productivity were rising. The differences, however, are probably not statistically significant.

If police productivity were increasing over this period at anything like the rate at which private sector productivity was increasing, it would be responsible for a large drop in crime. Of course, police may not be like the average private sector worker: the public sector may provide smaller rewards to innovation, reallocation of resources from less productive to more productive establishments is hard, and reallocation across sectors is by definition impossible. Increases in criminal productivity, moreover, may offset increases in police productivity. On the other hand, the federal government invested heavily in police-related research, starting in the 1970s, and much defense technology is transferable to policing. Police agencies were under considerable pressure to improve performance, and police executives under whom crime declined were rewarded well. Many police officers worked part-time in the private sector, and most prepared for a job in the private sector after retirement; they were not cut off from private sector advances. Television shows and movies continued to treat police officers as heroes. By how much could this sector lag behind the rest of the economy?

Leaded gasoline

As we saw in [Section 23.2](#), exposure to lead *in utero* and in childhood causes impulsiveness, aggressiveness, and low IQ in young adulthood, which plausibly cause violent crime. Lead in gasoline was phased out in the United States between 1975 and 1985. [Reyes \(2007\)](#) argues that this phaseout caused violent crime to fall in the 1990s. It had no measurable effect on murder and property crime.

This hypothesis is plausible and has some other implications that are borne out by the data. For instance, it implies bigger crime drops in areas where lead had been most pervasive in the atmosphere—large dense cities. Indeed, violent crime fell more in the largest cities than in less densely populated places. Thus, removal of lead from gasoline explains part of the residual in the violent crime drop of the 1990s.

How much it explains is a more complex issue. Reyes runs regressions that imply that changes in the lead content of gasoline caused a 56% decline in violent crime between 1992 and 2002 (almost the same period that Levitt uses)—more than the entire actual decline in violent crime. Taken at face value, the result means that powerful forces were at work in the 1990s to raise the violent crime rate, but we do not know what these forces were. It replaces one hard question (what moved violent crime down?) with an even harder question (what would have moved violent crime up were it not for the phaseout of leaded gasoline?).

The leaded gasoline hypothesis is like the abortion hypothesis in many ways. Events at the beginning of a cohort's life alter the distribution of criminogenic characteristics in that cohort; nothing happens for a decade or more; and then the effect bursts forth in lower crime rates when the cohort reaches crime-committing age. Reyes makes the analogy closer by using the same specification that [Donohue and Levitt \(2001\)](#) used and running it on almost the same set of years. (Reyes also includes the Donohue–Levitt “abortion rate” in her regressions, and finds that this variable accounts for an additional 29% violent crime decrease—but no significant change in murder or property crime.)

The variable that Reyes uses for a state-year is the lead content of the average gallon of gasoline sold. This depends on the proportion of premium gasoline sold, and on the oil companies that dominate in that state and their relationship with the Environmental Protection Agency. The actual exposure of poor children to lead from that gasoline depends on the number of gallons used, the proportion of young children in the population, and the distance between young children and traffic. For 4 years, 1976–1979, Reyes has data on lead in children's blood, and finds a correlation between that and the lead content of the average gallon of gasoline, her independent variable. But the relationship is not tight: even with several covariates and multiple fixed effects, R^2 is only 0.27 ([Reyes, 2007](#), table 3, column 3).

Like the abortion hypothesis, the leaded gasoline hypothesis implies that the crime decline should be a cohort phenomenon, while it was in fact a year phenomenon: arrest rates for all cohorts turn around at about the same time. Unlike the abortion hypothesis,

however, the leaded gasoline hypothesis is supported by some evidence that changes in lead exposure in the 1970s and 1980s altered behavior in dimensions other than crime. Reyes (2012) uses individual-level data to show that lead exposure *in utero* and in early childhood among cohorts born in the late 1970s and early 1980s led to increased behavioral problems in childhood and in adolescence.

The individual-level data, however, do not imply changes in behavioral problems as large as those that the state-level data implied for violent crime. For instance, the elasticity of “hitting and hurting someone” with respect to lead exposure is about half the elasticity of violent crime with respect to lead exposure. The leaded gasoline hypothesis thus probably explains a part of the violent crime residual. It is hard to evaluate because the relevant lead exposure has so far been approximated only roughly. For a variety of reasons having little to do with scientific merit, the lead hypothesis has attracted neither the attention nor the rigorous probing that the abortion hypothesis has had. That is unfortunate.

In utero exposure to tobacco and alcohol

We noted in Section 23.2.2.3 that people whose mothers smoked or drank alcohol during pregnancy were more likely to commit crimes, especially violent ones. Declines in maternal smoking and drinking might then cause crime declines 15–20 years later. This possibility has not been systematically investigated, to our knowledge, but it deserves serious consideration.

Smoking among pregnant women in the United States was once common. In 1964, in a large but not completely representative sample, nearly half of pregnant women in the third trimester smoked, with half a pack a day being the average consumption among smokers (Aizer and Stroud, 2011, pp. 10–11). We do not have a time series on smoking by pregnant women. Smoking by educated women began to fall almost immediately after the Surgeon General’s report on smoking in 1964, but less educated women did not react as quickly as more educated women did. The educational gradient in smoking became steeper between the 1960s and 1980, and so did the educational gradient in newborn health. After that, the gradients flattened, and returned to the steepness of the 1960s in the early twenty-first century. Aizer and Stroud (2011) interpret this record as an indication that information about the baleful effects of smoking diffused more slowly among less educated women than among more educated women.

For crime, the behavior of less educated women may be more relevant. The Aizer–Stroud record thus suggests that maternal smoking in this group may have started to fall in the 1970s, and thus like legalized abortion and the leaded gasoline phaseout, produced less criminogenic cohorts coming of age in the 1990s.

Drinking by pregnant women was similarly unmeasured during the 1970s and 1980s. Drinking in general, though, seems to have peaked in the late 1970s, and so if pregnant women mirrored the general population, cohorts coming of crime-committing age in the mid to late 1990s would have had fewer criminogenic characteristics. In both

government and private data, 1978 is the peak year for the proportion of the adult population who drank alcohol in the past month (U.S. Department of Health and Human Services, 1991; Newport, 2010). Fetal alcohol syndrome was identified and named in the medical literature in 1973 (Jones and Smith, 1973), and in 1981 the Surgeon General recommended that women abstain from alcohol while pregnant or planning a pregnancy. So awareness of the dangers of alcohol during pregnancy may have been rising before alcohol consumption in the general population peaked.

As with the other two cohort explanations (abortion and leaded gasoline), any attempt to demonstrate that changes in maternal smoking and drinking behavior caused part of the crime decline of the 1990s faces two hurdles. First, finding those changes will be hard. Second, cohort stories are not congruent with the almost simultaneous reduction in arrest rates for all age groups in the early 1990s.

Psychopharmaceuticals

As we saw in Section 23.2, greater use of psychopharmaceuticals causes a reduction in violent crime (Cuellar and Markowitz, 2007; Marcotte and Markowitz, 2011). During the 1990s, the prevalence of mental illness did not increase, but the proportion of mentally ill people receiving treatment grew by almost half (Marcotte and Markowitz, 2011). Most of the expansion of treatment used psychopharmaceuticals, and treatment in general shifted in that direction, with a number of new drugs coming on the market.

Marcotte and Markowitz find that “new-generation antidepressants” and stimulants reduced violent crime when they study the 1997–2004 period. The new-generation antidepressants trazodone and bupropion were approved by the Food and Drug Administration only in 1988 and 1985, respectively. They estimated that increased prescription of new-generation antidepressants was responsible for about 5% of the decline in violent crime between 1997 and 2004. If the same ratio held for 1991–2001, it would be responsible for about a 2 log point decrease.

Interactions

As discussed in Section 23.4.5, the distribution of individuals across space and social networks matters when peer effects can influence crime (Glaeser et al., 1996). A change in the proportion of violent or desperate individuals in a community, for instance, can induce changes in the behavior of others. These effects are highly nonlinear: for instance, an even distribution across space of those most prone to violence is likely to result in a very different rate of homicide than one where such persons are concentrated in a few locations (O’Flaherty and Sethi, 2010c). In this case, responses to the many large changes in crime-relevant variables would be nonlinear, and attempts to find linear effects for smaller changes would fail to predict what happened in the 1990s.

Private protection

Cook and MacDonald (2011, p. 333) write about the attempts to explain the crime decline of the 1990s: “A surprising feature of this speculation has been the absence of attention to the role of private actions to prevent and avoid crime.” Anderson (1999) finds that private spending on crime prevention is of the same magnitude as public spending, and with fewer principal-agent problems it may be more effective. As we saw in Section 23.4, the few private security initiatives that have been rigorously evaluated have been shown to be highly efficacious.

To the extent that private protection efforts have been measured, they expanded rapidly in the 1990s, and their productivity may have increased at least at the same rate as the average private sector activity. Cook and MacDonald (2011, p. 335–336) show that security guards were already more numerous than police officers at the start of the decade, and their employment grew at approximately the same rate. The proportion of crimes reported to the police also grew; reporting a crime is an indispensable form of aid that private parties must provide if perpetrators are to be apprehended.⁵⁹

Business Improvement Districts (BIDs) also became more numerous in the 1990s. Although the first BID was established in 1970 (Bloor Street in Toronto), the major growth in the United States appears to have started in the 1980s. New York City has a good consistent history of BIDs. The first BIDs were established in the 1970s, usually in connection with specific attempted malls. In the 1980s, they spread quickly. In all, 31 of New York City’s 67 BIDs were founded between 1980 and 1995. Weighting by size would make this concentration even more apparent. Of the four largest BIDs, those with annual assessments over \$9 million now, three were founded in the early 1980s (34th Street, Times Square Alliance, Grand Central Partnership), and one (the Downtown Alliance) was founded in 1988. The Furman Center for Real Estate and Urban Policy (2007) also suggests that the larger BIDs are more effective at reducing crime, since many of the smaller ones have almost no budget for security.

Houses also became safer, in part through the spread of gated communities (Helsley and Strange, 1999). The American Housing Survey first collected information on “secure communities” in 2001; it found 7 million households in these communities. A more restrictive class in the American Housing Survey, secure communities with special entry systems, contained 4 million households. The number of people living in gated communities, and the proportion of people living in these communities, probably rose in the 1990s. So did the proportion of households with garages or carports—a deterrent to motor vehicle theft—even as the vehicles themselves became harder to steal.

⁵⁹ The rise in the reporting rate, however, may be a figment of selection. Within any category of crime, the more serious offenses are more likely to be reported. If crime fell selectively, with greater decreases in less serious crimes than in more serious ones, then the reporting rate would go up with no change in the public disposition to report crime. O’Flaherty and Sethi (2008) show that for robberies, the average seriousness increased as the number fell.

Burglar alarms became cheaper, more effective, and more popular: the proportion of homes with alarm systems rose from 1% in 1975 to 25% in 2003 (Lee, 2008). Surveillance cameras also proliferated in the 1990s. These devices, more properly known as closed-circuit television (CCTV) systems, became practical in the 1970s when videocassette technology allowed pictures to be stored easily and recovered later; before that, CCTV required constant monitoring. The major technological breakthrough of the 1990s was the digital multiplexer, which allowed motion recording and permitted the output from several cameras to be recorded at the same time and stored cheaply and accessibly. The weight of research indicates that CCTV systems probably reduce property crime and possibly robbery, both to a modest degree; see Welsh and Farrington (2009) for a meta-analysis.

Electronic article surveillance—the tags placed on clothing and other merchandise that activate alarms or spill ink if they are removed from the store—also developed at about the same time. The first systems appeared in the 1970s, major technical advances were first made in the 1980s, and the technology has continued to improve. While these devices may have reduced actual shoplifting, their effect on reported shoplifting is ambiguous, since they cause a higher proportion of thefts to be detected.

Electronic banking

The 1990s also saw the dissemination of electronic banking technologies. Although ATMs were invented in the late 1960s, there were only 80,200 terminals processing 5.75 billion transactions in 1990. By 1999 there were 227,000 terminals and 10.89 billion transactions (Statistical Abstract, 2001, table 820). ATMs allow consumers to make smaller and more frequent cash withdrawals, and allow businesses to make more frequent cash deposits. Thus, both consumers and businesses can carry less cash and will be less inviting targets for robbery, larceny, and burglary.

Other banking innovations allowed businesses and consumers to dispense with cash entirely. The proportion of households with at least one general-purpose credit card rose from 56% in 1989 to 73% in 2002 (Statistical Abstract, 2012, table 1189). Debit cards were used sparingly before 1990, but the number of point-of-sale terminals rose from 53,000 in 1990 to 2.35 million in 1999, and the number of point-of-sale transactions rose from 191 million in 1990 to 2.48 billion in 1999 (Statistical Abstract, 2000, table 820). The proportion of households using debit cards rose from 20% in 1995 to 50% in 2001 (Statistical Abstract, 2012, table 1185). Direct deposit also grew: 53% of households used direct deposit in 1995 and 71% used it in 2001 (Statistical Abstract, 2012, table 1185).

Electronic banking also affected how governments disbursed benefits, especially to low-income, often unbanked households who were disproportionately victims of robbery and larceny. In the 1990s, the primary focus was on food stamps, a large program that gave low-income households a monthly allotment of paper stamps that they could exchange for food (and only food) at the stores of participating retailers. Food stamps

were essentially cash, and could be exchanged reasonably easily in black markets. They were worth stealing. During the 1990s, most states, with federal assistance and urging, replaced paper food stamps with “electronic benefits transfer” (EBT) (Pirog and Johnson, 2008). Under EBT, recipients use debit cards or smart cards, rather than stamps, to purchase food; they do not need bank accounts, and the cards are PIN-protected.

Other transfer programs lagged behind food stamps, but still reduced their reliance on paper. Temporary Assistance for Needy Families provides income on a monthly basis for eligible families; in many states checks were all delivered on the same day, and recipients had to negotiate them and hold cash for the rest of the month. By 2003, 33 states delivered Temporary Assistance for Needy Families benefits by debit or smart cards, and a few other states offered a direct deposit option. By 2001, 13 states delivered general assistance by EBT, 12 states delivered refugee assistance, and nine states delivered Supplemental Security Income (Stegman et al., 2003).

Social security payment systems also changed during the 1990s. The proportion of families using direct deposit was 50% in 1990; it was 75% in December 1998, and probably rose another several points by 2001. The Social Security Administration also started to spread the day of the month on which recipients received their benefits. Before 1997, all benefits were issued in the first week of the month; beginning in that year, the benefit days for new recipients were spread throughout the month (U.S. Social Security Administration, 2000, chapter 5).

The effect of this revolution on crime has been little studied. The best recent article is that by Wright et al. (2014), which examines the phase-in of EBT for food stamps and welfare payments in Missouri. EBT came to different counties at different times, and this variation allows the study authors to identify changes in crime. EBT reduced assault, burglary, and larceny by about 10% in the average (population-weighted geometric mean) county. It did not have a significant effect on robbery, rape, or motor vehicle theft.

The results on assault and robbery are perhaps surprising, since robbery is an acquisitive crime and assault is not. Robbery in Missouri is heavily concentrated in a few counties, and so the study authors believe that they do not have enough variation to see an effect at the county level. Why EBT should affect assault is a harder question. Perhaps the presence of cash leads to more frequent and more serious disputes, but we do not know what kind of disputes these are (e.g., whether they are over drugs or criminal activities or love or domestic responsibilities).

Online shopping

Amazon customers cannot shoplift. Nor are they likely to be accosted while walking to the store, or to have their cars broken into in the store parking lot, or to leave their homes and allow a burglar to enter. Internet commerce is far from crime-free, but it probably reduces the exposure of both merchants and customers to traditional index crimes.

Retail e-commerce rose from essentially nothing in 1991 to \$34.6 billion in 2001 ([Statistical Abstract](#), 2012, table 1055).

In all, during the 1990s people's lives changed in many ways that reduced index crime. Better cars, more security guards, BIDs, stronger houses, more garages, securer communities, burglar alarms, surveillance cameras, ATMs, credit and debit cards, direct deposit, EBT, online shopping, and cell phones—in all these ways safety increased. Property crime was probably affected the most, but robbery, a violent crime, was also affected; [O'Flaherty and Sethi \(2009a\)](#) infer from rates of resistance and violence that “victim hardening” was part of the story of declining robbery, but the size of the contribution remains to be estimated.

23.6.2.3 *Wrapping up: the great American crime decline of 1991–2001*

After the traditional explanations were accounted for, we found yet-to-be-explained residuals of about 30 log points for murder, 20 for violent crime, and 10 for property crime. Unobserved net increases in police productivity should operate across the board, changes in how potential offenders grow up and their minds operate should affect mainly violent crime, changes in how potential victims live should affect mainly property crime, and interaction effects should affect mainly murder. The fall in murder seems to have the least explanation, but improvements in emergency medical treatment may play a role (however, offenders may offset these technological improvements with improvements of their own, like semiautomatic weapons and high-caliber weapons). We probably could explain more of the murder decline too if we could explain why the crack epidemic happened in the 1980s and nothing like it recurred in the 1990s. The great American crime decline of the 1990s is not a mystery; we just do not know enough about the non-traditional explanations yet to tell the whole story.

23.6.3 The modest American crime decline of the twenty-first century

The American crime decline continued into the first decade of the twenty-first century, although the rate of decline slowed. While all four traditional explanatory variables moved in the direction of lower crime in the 1990s, only imprisonment moved in that direction in the 2000s, and the movement of imprisonment was small. Hence, the residuals did not shrink much.

For comparability, we will continue to use the format [Levitt \(2004\)](#) adopted and follow murder, violent crime, and property crime for the period 2001–2011. During this period, UCR murder fell by 17.5 log points, violent crime fell by 26.5 log points, and property crime fell by 23 log points (within that category, motor vehicle theft fell by 62.7 log points).

On demography, the proportion of residents who were male and between 15 and 25 years and between 15 and 30 years did not change between 2000 and 2010; so demography plays essentially no role in this decade. Per capita police fell by 2.9 log points, from

2.45 to 2.38. This should have increased crime slightly. The general unemployment rate rose from 4.2% in 2001 to 9.1% in 2011; this should also have increased property crime. The imprisonment rate rose only slightly, by 3.5 log points.

Table 23.4 redoes Table 23.3 for the period 2001–2011, following the same parameters. (“LP” means we use parameters from Levitt (2004), since he did not follow this decade.)

For murder and property crime, the traditional variables explain essentially none of the crime decline; it does not matter which parameters are used. The traditional variables project an increase in property crime, with the Levitt parameters projecting a smaller increase. In either case, the message is the same: the 2001–2011 crime decrease was entirely (or more than entirely) in the residual. The violent crime residual is about the same as the residual we estimated for the 1990s; the murder residual is smaller, and the property crime residual is larger.

This result is encouraging for the speculation about the residual in Section 23.6.2.2, because many of the processes we thought might be producing the residual in the 1990s continued or intensified in the 2000s.

The data that Marcotte and Markowitz (2011) employ on the increased use of psychopharmaceuticals, for instance, extend to 2004. Labor productivity grew in the non-farm business sector at about the same rate (2.7% per year from 2000 to 2007, 1.9% per year from 2007 to 2012; versus 2.2% for 1990–2000), and defense and antiterrorism research and investment were significant; hence, there is no reason to expect a fall-off in police productivity growth. BID growth in New York City slowed but did not stop, and the rest of the country may have lagged behind New York.

Housing continued to become securer. Surveillance cameras continued to spread, spurred in part by concern over terrorism; newspaper stories in 2013 about surveillance cameras in New York City apartment buildings asked not whether they would have them, but how many they would have (Kaufman, 2013).

Table 23.4 Estimates of log point changes in crime due to changes in traditional variables, 2001–2011

	Murder			Violent crime			Property crime		
	LP	Smaller	Bigger	LP	Smaller	Bigger	LP	Smaller	Bigger
Police size	+1		+2	+1		+2	+1	+0.5	
Prison	−1	−0.5		−1	−0.5		−1	−0.5	
Demography	0			0		0	0		0
Macroeconomy	0			0			+5		+10.5
Total	0	+0.5	+1	0	+5	+1	+5	+5	+10.5
Actual (UCR)	−17.5	−17.5	−17.5	−26.5	−26.5	−26.5	−23	−23	−23
Residual	−17.5	−18	−18.5	−26.5	−27	−27.5	−28	−28	−33.5

LP, parameters from Levitt (2004).

ATM and debit card usage continued to grow. From 2001 to 2007, the proportion of households using ATM cards rose from 57% to 76%, and the proportion using debit cards rose from 50% to 71% ([Statistical Abstract](#), 2012, table 1185). The volume of purchases with debit cards rose from \$311 billion in 2000 to \$1.45 trillion in 2009 ([Statistical Abstract](#), 2012, table 1187).

On the other hand, several processes that may have contributed to the residual in the 1990s were probably less important. There were 23 fewer executions in 2011 than in 2001; on the maximum estimates of the deterrent effect of capital punishment, this would have increased murders by 1 log point. Security guard employment grew, but not as fast as the population, and the number of security guards per capita fell by 5 log points ([U.S. Bureau of Labor Statistics](#), 2013b). The cohort effects from leaded gasoline and *in utero* exposure to tobacco and alcohol were still operating in the early 2000s, but not as strongly. Burglar alarms probably did not spread as quickly. Direct deposit reached 71% of households in 2001, and this grew only to 80% by 2007 ([Statistical Abstract](#), 2012, table 1145). Although California was one of the few states that was late to switch to food stamp EBT, most (but not all) of the conversions took place before the turn of the century. (Social security added a debit card in 2008 to convert unbanked stragglers away from paper checks.) Similarly, credit card usage grew, but the rate of growth tapered off.

But other trends that reduced crime strengthened in the twenty-first century. Cars became much harder to steal, and older, easier-to-steal models disappeared from the streets. Online shopping volume rose from \$34.6 billion in 2001 to \$194 billion in 2011. Cell phone subscriptions rose from 128.4 million in 2001 to 300.5 million in 2011. Sophistication also grew. Cell phones became ubiquitous surveillance cameras, with everyone ready to take a picture anywhere any time. They also became sources of information; fewer people became lost, and even when they were lost they did not appear to be so. On the other hand, smartphones became attractive targets for robbery and theft, although technological solutions to this problem are likely before this chapter is published ([Chen](#), 2014).

Online shopping may also have extended to vice, most notably pornography and prostitution, probably with gains in security. (Prostitutes who do not solicit in public may still be victims of pathological murderers, but they and their customers are less likely to be victims of petty robbers.) It is possible that this trend has also affected the anonymous sale of illicit drugs. For instance, according to a federal indictment, the website Silk Road operated between 2011 and 2013 and “was used by several thousand drug dealers and other unlawful vendors to distribute hundreds of kilograms of illegal drugs and other illicit goods and services to well over a hundred thousand buyers” ([U.S. District Court, Southern District of New York](#), 2013). [Zimring \(2011\)](#) argues that open-air drug markets have disappeared from New York City, but drug consumption has not fallen.

The spread of the Internet per se may also have reduced index crime. To the extent that young men spent more time at home playing video games and chatting with friends,

they may have spent less time outside where they could be either victims or offenders. Rational acquisitive criminals may have decided that the expected return from running online scams exceeded the expected return from stealing pocketbooks, and so redirected their attention. Cybercrimes are not index crimes, and the data show only that index crimes decreased in this decade.

Thus the residuals in the first decade of the twenty-first century are similar in size to the residuals in the 1990s, and so are the candidates to explain them.

23.6.4 Developed countries in the twenty-first century

Most of the candidates for explaining the residuals in the United States are varieties of technological change, either currently (e.g., electronic banking) or in the recent past (e.g., leaded gasoline), and so one would expect that other developed countries would experience similar trends. This does not imply that crime should actually fall in all developed countries, since the traditional factors can move in different directions; it implies only that the residuals should generally be negative.

As we saw in [Table 23.2](#), the overwhelming direction of change in the four best-reported crimes has recently been negative in most OECD countries. This is despite the Great Recession and austerity programs, which have put pressure on police and prison budgets. Greece is an outlier: not surprisingly, crime rose in all categories there. Otherwise, the trend is clear and consistent. Aside from Greece, homicide went down in 21 of 23 countries, robbery went down in 15 of 21 countries, burglary went down in 17 of 18 countries, and motor vehicle theft went down in 18 of 18 countries. The median change (including Greece) was a decrease larger than the US decrease in all crimes but robbery. The huge decrease in motor vehicle theft in high-income countries is especially strong evidence for technological change.

We have not calculated residuals for all these countries, but doing so would be a valuable exercise.

23.6.5 The great American crime rise, 1965–1975

Between 1965 and 1975, reported index crime in the United States basically doubled. Part of this increase probably came from better reporting (murder and motor vehicle theft showed slightly smaller increases than the other index crimes). As [Table 23.5](#) shows, the rise was general throughout the country, although somewhat smaller in the South.

The rise also affected all types of communities, although it was greater in large (and medium-sized) cities. Sometimes the rise is thought of as occurring primarily among African Americans, but to the extent that arrests reflect actual offending, this does not appear to be the case, except possibly for motor vehicle theft. [Table 23.6](#) shows the proportion of arrestees who were black for the various index crimes in 1966 and 1975. For the more serious crimes, the proportion generally declines by small amounts, and only for

Table 23.5 Change in murder and motor vehicle theft rates, by region, 1965–1975

	Murders per capita			Motor vehicle thefts per capita		
	1965	1975	1975/1965	1965	1975	1975/1965
New England	3.6	7.6	1.11	299.4	652.8	1.18
Mid Atlantic	4	8.9	1.23	282.6	534.1	0.89
North Central	3.7	8.1	1.19	244.5	431.4	0.76
West North Central	3.1	5.5	0.77	176.4	328	0.86
South	8	12.7	0.59	175.7	329.8	0.88
East South Central	8.4	12.7	0.51	130.6	273.3	1.09
West South Central	7	12.4	0.77	178.5	351	0.97
West	4.2	9	1.14	351.5	539.1	0.53
Pacific	4.3	9.4	1.19	388.3	585	0.51
United States	5.1	9.6	0.89	254.4	469.4	0.85

Sources: UCR, 1966, 1975.

Table 23.6 Proportion of arrestees who were black, 1966 and 1975

	1966	1975
Murder	57.2%	54.4%
Rape	46.8%	45.4%
Robbery	57.7%	58.8%
Aggravated assault	48.9%	39.5%
Burglary	31.4%	28.4%
Theft	29.7%	30.6%
Motor vehicle theft	19.6%	26.4%

Sources: UCR 1966, 1975.

motor vehicle theft was there a substantial increase. Crime by whites almost certainly doubled in this decade too.

The traditional variables do not explain this crime rise. Police officers per 10,000 population rose from 1.7 to 2.1 (UCR), and prisoners per 100,000 rose from 108 to 111. These two changes should have decreased crime. On the other hand, unemployment rose from 4.6% in June 1965 to 8.8% in June 1975, which should have increased property crime, and much of the baby boom cohort came of crime-committing age. [Zimring \(2007\)](#), in analyzing a slightly different period (1960–1970) offers an estimate that changes in age structure in that period would have increased index crime 13%. (We will use this figure for our table.)

[Table 23.7](#) shows Levitt-style estimates for the traditional factors for the 1965–1975 period (along with Zimring's demographic estimate). The residuals are huge, and call for an explanation. Lead gasoline is one possible factor. After World War II, driving increased and so did the lead content of gasoline. Cohorts who came of age in the 1960s and 1970s suffered from increasing exposure to lead *in utero* and in childhood.

Table 23.7 Estimates of log point changes in crime due to changes in traditional variables, 1965–1975

	Murder	Violent crime	Property crime
Police size	−8	−8	−8
Prison	−1	−1	−0.5
Demography	+13	+13	+13
Macroeconomy	0	0	+4
Total	+4	+4	+8.5
Actual change (UCR)	+63	+89	+76
Residual	+59	+85	+67.5

One strength of the lead hypothesis for the 1990s is that it also says something about the 1960s and 2000s.

Aside from lead, no obvious hypotheses present themselves. [Miron \(1999\)](#) argues that the War on Drugs, which was declared in 1971, was responsible for much of the rise, but the argument has not gained traction. A popular view is that crime rose among African Americans as a natural continuation of the riots of the 1960s, but whites did not riot in the 1960s and crime rose among them in parallel. The combination of a draft and an unpopular war may have reduced the threat of prison in the mid 1960s—how much worse was San Quentin than Khe Sanh?—but the draft was over by 1972. The completion of the interstate highway system and the expansion of air travel may have increased mobility and productivity among criminals with no corresponding gains in law enforcement productivity, but no serious research has been done on this topic.

Becker's seminal article that launched the modern economics of crime was published in the midst of the great crime rise. Looking at residuals from this period is a good way of understanding why criminologists of that era were not impressed by what economists were doing. The economics of crime at this time was really about arranging the deck chairs on the Titanic. But no one has identified the iceberg yet. The great American crime rise remains a mystery, and a great topic for research.

23.7. WHERE ARE CRIMES COMMITTED?

Just as the traditional factors account for little of the interesting variation in index crime over time, they also seem to account for little of the interesting variation in crime over space. This is not to argue that a jurisdiction that abolished its police department or adopted capital punishment for jaywalking would look the same as the rest of the world. There just are no such jurisdictions.

23.7.1 Intermetropolitan variation

Index crime, for the most part, is a nontradable activity and so we do not expect to see specialization by metropolitan area. (Exceptions may be tied to the import of illicit drugs,

Table 23.8 Geographic concentration indices for crime and other phenomena

	States	Intrametropolitan (Newark)
Murder	0.0044	0.262
Rape	0.0031	0.073
Robbery	0.0028	0.181
Aggravated assault	0.0024	0.111
Burglary	0.0028	0.033
Larceny-theft	0.0013	0.010
Motor vehicle theft	0.0156	0.188
Poverty (individuals)		0.077
Black population		0.095
Automobile manufacturing	0.127	
Automobile parts	0.089	
Photographic equip.	0.174	
Carpet manufacturing	0.378	
Soft drinks	0.005	
Manufactured ice	0.012	
Newspaper	0.002	
Miscellaneous concrete	0.012	

but this is indirect; and for reasons of evasion, the dispersion of import facilities may be greater than that for legal and bulky commodities such automobiles.)

We computed the Ellison–Glaeser concentration index for each index crime at the state level from the 2012 UCR. The values were in the same range as those for industries that [Ellison and Glaeser \(1997, p. 902\)](#) described as “industries that one could not imagine to be concentrated. . .the bottled and canned soft drink (SIC 2086), manufactured ice (SIC 2096), newspaper (SIC 2711), and miscellaneous concrete products (SIC 3272) industries.” [Table 23.8](#) provides more detail.

In the United States, crime is only weakly associated with larger metropolitan areas. Following Glaeser and Sacerdote (1999), we regressed the logarithm of crime per capita (from the 2012 UCR) on the logarithm of population and a vector of region dummies (nine regions). [Table 23.9](#) reports the results.

Rape, burglary, and larceny do not appear to be correlated with population size at all, while the elasticity of aggravated assault is small. But robbery, motor vehicle theft, and murder are considerably more prevalent in larger metropolitan areas.

In Canada, the relationship is weaker. Canada uses a “crime severity index,” which weights crimes by average sentence. The elasticity of the violent crime severity index with respect to the population is an insignificant 0.05, and the elasticity of the total crime severity index is an insignificant -0.03 ; the data are from [Perreault \(2013, table 4\)](#).

Why robbery and motor vehicle theft should be strongly correlated with metropolitan area size in the United States is an open question that has not been investigated. For

Table 23.9 Elasticities of crime rates with respect to the population: US metropolitan statistical areas, 2012

Murder	0.16★ (5.01)
Rape	−0.02 (−0.79)
Robbery	0.33★ (13.28)
Aggravated assault	0.08★ (2.92)
Burglary	0.01 (0.26)
Larceny-theft	0.01 (0.55)
Motor vehicle theft	0.23★ (10.90)

Region fixed effects.

Number of observations: 353. *t* values are given in parentheses.

★Significant at the 5% level.

Source: FBI, Crime in the United States, 2012, table 6.

both crimes, the search for a suitable victim may be a large part of the cost of committing the crime, and larger (and denser) metropolitan areas may be more attractive places to search. Murder and assault may be more prevalent in larger metropolitan areas because of historically higher levels of atmospheric lead, or because encounters between people that could escalate into conflicts are commoner.

23.7.2 Intrametropolitan variation

Intrametropolitan concentration of crime seems to be greater than intermetropolitan concentration.

We use the New Jersey part of the Newark metropolitan area as an example.⁶⁰ Six municipalities that have reputations as “rough towns” account for 24.9% of the metropolitan area population and occupy 2.6% of the land, but in 2011 they were the location of 85.6% of murders, 79.9% of robberies, 77.4% of motor vehicle thefts, and 73.4% of aggravated assaults. Burglary and larceny were not so heavily concentrated: only 51.8% of burglaries and 34.9% of thefts were reported in the six towns.

⁶⁰ The subparts of the Newark metropolitan area are Newark, East Orange, Irvington, Orange, the balance of Essex County; Elizabeth, Plainfield, the balance of Union County; Hunterdon County; Morris County; Somerset County; and Sussex County. We omit Pike County, Pennsylvania, which represents less than 3% of the metropolitan area.

More formally, we computed Ellison–Glaeser concentration indices for crimes in the Newark metropolitan area. These are shown in [Table 23.8](#).⁶¹ Intrametropolitan concentration, especially for murder, robbery, and motor vehicle theft, is much greater than intermetropolitan concentration. We also computed concentration indices for poverty and for the African American population. Murder, robbery, and motor vehicle theft are more concentrated than these residential characteristics, but burglary and theft are not. (Aggravated assault is slightly more concentrated.)

Theft is by far the most numerous index crime—burglary is the second most numerous—and is also the least concentrated geographically. Hence, analyses that focus on aggregate index crimes miss a great deal of concentration.

It is important to realize that these data are about where crimes are committed (or, in the case of murder, where bodies are found), not where criminals or victims live. In a metropolitan area, people cross municipal and neighborhood boundaries easily and often. Paramus, New Jersey, for instance, has one of the highest rates of theft in the state—not because its residents are disproportionately felonious or disproportionately easy marks, but because it is the home of several large malls. People go to Paramus from throughout the metropolitan area to shop, to work, to steal, and to be stolen from. Similarly murder in Newark is not restricted to Newark residents, on either side of the gun.

Finer partitions of geography also show great concentrations of crime. Within a city, a small number of “hot spots”—intersections or addresses—are sites of a disproportionate amount of crime, especially robbery and motor vehicle theft. [Sherman et al. \(1989\)](#) examined police calls for service during a year in Minneapolis. If calls were Poisson distributed with equal probability in all places, only 6854 places would have had no calls, and almost none would have had more than 14. But in fact, 45,561 places had no calls, and 3841 had more than 15 calls. Those 3841 places—3.3% of all the places in the city—generated 50.4% of all calls. Similarly, independent Poisson distributions imply that almost no place should have seen more than two robbery calls or three motor vehicle theft calls, but 293 places had more than two robbery calls (one place had 28) and 541 places had more than three motor vehicle theft calls (one place had 33) ([Sherman et al., 1989](#), table 2).

⁶¹ All measurements in [Table 23.8](#) use the Ellison–Glaeser (1997) index of geographic concentration. For crimes and population characteristics, the index is G_i ; for industry employment the index is γ (see their paper for definitions). Crime data reflect the concentration of reported crime relative to the total population. Crime and population data for states are from the 2012 UCR. Crime and population data for subparts of Newark metropolitan area are from the 2011 annual report of New Jersey State Police. Poverty data reflect people in poverty relative to people for whom poverty status is determined. The data source is the 2010–2012 American Community Survey, except for Irvington and Orange, for which the data source is the 2007–2009 American Community Survey. Race data reflect people who identify themselves as a member of one race, black or African American, relative to the total population. The data are from the 2007–2009 American Community Survey. Industry data reflect the concentration of employment relative to total employment. These data are from [Ellison and Glaeser \(1997, p. 902\)](#).

What makes particular locations magnets for crime is a question that has received considerable attention. The traditional explanations for temporal variation probably cannot explain a large proportion of intrametropolitan spatial variation (poverty is a possible exception, but we have already seen that serious crimes are more spatially concentrated than poverty). In the Newark metropolitan area, the towns with a high crime rate are not conspicuous as magnets for young men, the prison system is the same for the entire metropolitan area (as we have defined it), and police strength works in the opposite direction (with 24.9% of the population, the six towns employ 35% of the metropolitan area's non-civilian police employees).

23.7.3 Explaining intrametropolitan variation

The analysis of [Sherman et al. \(1989\)](#) suggests that crime ought to be viewed as being concentrated at *places*—defined as intersections or addresses—rather than neighborhoods. Even in high-crime neighborhoods the vast majority of places are crime-free over substantial periods. Given the ability of victims to respond to criminal activity through avoidance or other precautions, this extreme concentration at places appears puzzling. For instance, consider the Minneapolis bar with the highest raw frequency of calls to police for predatory crimes in their data: “With 25 robberies in 1 year, and an estimated mean daily population of no more than 300, Moby Dick’s Bar had a robbery call rate of 83 per 1000 persons—7 times higher than the call rate of 12 per 1000 for the city’s entire 1986 estimated population of 362,000” ([Sherman et al., 1989](#), p. 44).

It is important to bear in mind that this concentration is stable over time, and occurs even in the face of any adjustments in behavior by victims, police, or local business owners. Since potential victims tend to avoid places with a reputation for violent crime, and law enforcement officials tend to divert resources to such locations, the intrametropolitan concentration of crime would be greater—perhaps substantially greater—in the absence of these mitigating effects. Accordingly, there must exist powerful forces that give rise to concentration.

[Sherman et al. \(1989\)](#) argue that some places—bars, liquor stores, adult theaters, or poorly lit parks—are *generators* rather than simply *receptors* of crime. Once establishments such as bars are in place, they cannot easily be moved in response to changes in the incidence of crime, nor can they easily alter the pool of individuals who congregate there. This can allow concentration to persist even in the face of adjustment by potential victims. Incapacitation of such places through better enforcement then need not lead to much displacement to other locations.

While certain types of places may indeed be potential generators of crime, there remains the question of why they act that way in some neighborhoods but not in others. Bars and liquor stores may be over-represented on the list of high-crime establishments in a metropolitan area, but most bars and liquor stores are not high-crime establishments.

The neighborhood in which the establishments are located clearly matters. In particular, concentrated poverty and residential instability are known correlates of violent crime at the neighborhood level. Sampson et al. (1997) argue that this link operates through *collective efficacy*, understood as a combination of social cohesion and informal social control; these ideas were anticipated in Jacobs (1961). Social cohesion refers to shared values and mutual trust in a neighborhood, while informal social control refers to a willingness to interfere in the face of local activities that are potentially damaging to the neighborhood if left unchecked—painting graffiti, loitering on corners, and physical altercations, for instance. The study authors use survey evidence to measure collective efficacy at the level of Chicago neighborhood clusters, and find it to be strongly correlated with measures of violent crime. They argue that the well-established association between crime and neighborhood characteristics such as poverty and residential instability operates in part through the channel of collective efficacy. They do not test for causality.

Understanding the spatial distribution of index crimes within cities requires us to broaden our focus to consider street vice—prostitution, illegal gambling, and drug selling. These are transactions with diffuse demand that spans geographic areas, income levels, and identity groups, but highly concentrated supply, much of it in central city neighborhoods. Well-developed theories of spatial competition descended from Hotelling (1929) and Salop (1979) can shed light on the concentration of street vice, and this in turn can help us understand the geographic distribution of index crimes. We consider here the case of drug selling, although similar arguments apply to other categories of street vice.

The sale of drugs involves fixed costs of protection, from rivals as well as from authorities. Sellers also use relatively unskilled, low-wage labor. Buyers are thinly spread over a large area and face significant risks when transporting large amounts of cash or contraband. Accordingly, they tend to make frequent small purchases. Under these conditions, the Salop (1979) model of spatial competition implies that areas with greater demand density will have more sellers per unit distance, as well as lower prices. Even if demand per unit population is uniform across locations, greater population density implies greater demand density, and hence inner cities will have lower prices and higher seller density than more sparsely populated suburbs. If trade between the city and the suburb is possible, competition from the city lowers suburban prices and reduces suburban seller density. If this effect is sufficiently strong, then all trade moves to the city, raising seller density and lowering prices there (O’Flaherty and Sethi, 2010a).

These considerations suggest that street vice will be most prevalent in centralized locations with high population density and low wages. But they do not account for the fact that, at least in the United States, street vice is most prevalent in neighborhoods that are predominantly black. This is not a recent phenomenon. A century ago, Booker T. Washington (1915) spoke of vice being openly paraded in black neighborhoods, next to schools and churches, attracting white customers from well beyond the immediate

vicinity. Decades later, [Myrdal \(1944\)](#) described the concentration of speakeasies in black neighborhoods during the prohibition era, along with gambling dens, cabarets, and the sale of narcotics, all catering to a clientele that included whites from far-flung locations. This racial character of the geography of vice requires explanation.

Unlike ordinary retail establishments, street vice imposes significant costs on local nonusers. Drug sellers and their customers are attractive targets for robbers, since they are likely to be carrying cash or valuable contraband, and are not inclined to approach law enforcement officials if victimized. Disputes between sellers cannot be settled in courts, so the threat of violence is pervasive. The marginal penalties associated with killing are lower for those already engaged in an activity that carries heavy sentences. This makes drug sellers more likely to kill, and for reasons discussed in [Section 23.4.5](#), also makes them more likely to be killed preemptively. Hence, the rates of robbery and homicide are liable to be high where street vice is prevalent. Illegal markets, in general, and open-air markets, in particular, are “high risk settings for violence” ([Reiss and Roth, 1993](#), p. 18).

These negative externalities make departure from afflicted neighborhoods attractive for those who can afford to leave, at least among the population of nonusers. Vice will be prevalent in neighborhoods with low incomes and low property values. Even modest preferences over neighborhood racial composition can then lead to the exodus from such neighborhoods of low-income whites. This process is self-reinforcing: if most whites leave, then the few who remain will be stereotyped as buyers from outside the neighborhood, and may themselves be victimized more frequently than blacks of similar income. As a result, street vice can come to be correlated with race, not because vice originates in or moves to black neighborhoods, but because its presence sets in motion movements of people that transform the racial character of residential locations ([O’Flaherty and Sethi, 2010a](#)).⁶²

As with any retail operation, street vice requires coordinated expectations between buyers and sellers regarding what can be found at what prices and at what locations. Two street corners that are otherwise quite similar may come to be very different over time with respect to their respective customer pools. More generally, locations will differ with respect to how lucrative they are for those who occupy them. The most desirable locations will tend to be occupied by the best protected sellers, who can fend off attempts at displacement. With well-settled expectations regarding turf, drug selling can proceed without frequent or extreme violence. But a disruption in these expectations can cause homicide levels to spike.

Sometimes such disruptions can come from the activities of law enforcement agents. If the police successfully target the most active and lucrative location and incapacitate the

⁶² This dynamic reinforces the segregating effect of racial stereotypes discussed in [Section 23.4.3](#); whites exit from high-crime neighborhoods more readily than blacks of comparable income because the former are stereotyped as being more compliant, and are therefore targeted by robbers at greater rates.

incumbent sellers there, competition for the vacated spots can result in violence. Because of this, a better strategy for law enforcement might be to target the *least* lucrative locations first, and ratchet upward toward more active and better protected areas. This strategy may take longer to have an impact on drug sales, but is less likely to precipitate a spike in homicide levels.

Evidence for the importance of this effect is provided by Dell (2012), who uses the outcomes of close municipal elections in Mexico as an identification strategy. In jurisdictions that narrowly elected mayors belonging to the conservative PAN party, enforcement actions increased significantly (relative to jurisdictions in which the PAN party candidate narrowly lost). The proposed mechanism was greater cooperation between local authorities and the federal administration of Felipe Calderón of the PAN party over the 2006–2012 period. Greater enforcement at these locations diverted drug trafficking to less heavily monitored routes to the lucrative US market. Working with a network model that predicts these alternative pathways, Dell finds significant increases in violence at these newly contested locations, and plausibly interprets these increases as arising from competition among rival organizations for “control of territories after crack-downs. . . have weakened the incumbent traffickers.”

23.7.4 Interjurisdiction variation

Economists have also looked at whether crime is greater in larger jurisdictions (it is) and why. In exploring this question, Glaeser and Sacerdote (1999) concentrated on total index crime, a number dominated by theft and burglary, and we have seen that spatial variation in these crimes differs from spatial variation in the other index crimes. They also use crime data for 1982, after the great American crime rise occurred but before the great American crime decline began, and that crime decline was especially strong in the two largest jurisdictions, New York City and Los Angeles. (Some high-crime cities such as Detroit and Philadelphia have also lost population since 1982, while the population of some low-crime cities has grown.)

In Table 23.10, therefore, we roughly redo the Glaeser and Sacerdote analysis with 2012 UCR crime data. (They restricted their sample to cities with a population over 25,000 for which complete information was available in the City and County Data Book, but the City and County Data Book was no longer published in 2012.)

The elasticities are somewhat larger than those for metropolitan areas (except for murder, where the exclusion of jurisdictions with zero murders may be affecting the coefficient), but follow the same basic pattern: location matters more for robbery and motor vehicle theft, less for larceny-theft and burglary. Glaeser and Sacerdote offer some theories on why these elasticities should be positive, and test them. The most powerful variable for explaining the positive elasticity of total index crime is the proportion of mothers without partners living in the jurisdiction. Something about larger jurisdictions

Table 23.10 Elasticities of crime rates with respect to the population: US police jurisdictions with over 25,000 people, 2012

	Observations	Elasticities
Murder	951	0.07★ (2.25)
Rape	1474	0.14★ (5.76)
Robbery	1571	0.59★ (18.67)
Aggravated assault	1578	0.38★ (13.27)
Burglary	1580	0.20★ (9.01)
Larceny-theft	1581	0.10★ (7.08)
Motor vehicle theft	1581	0.43★ (14.82)

Region fixed effects.

The number of observations differs because of zeroes and missing values. *t* values are given in parentheses.

*Significant at the 5% level.

Source: FBI, Crime in the United States, 2012, table 6.

attracts (or propagates) single mothers (or drives married mothers out), and single mothers attract (or propagate) crime.

What is missing from this analysis is an economic theory of jurisdictions. The theories that Glaeser and Sacerdote offer are in fact theories about metropolitan areas. There are several good economic theories about what determines the size of metropolitan areas, and how big metropolitan areas differ from small ones. There are no comparable theories about police jurisdictions. This makes it hard to interpret their results.

23.8. CONCLUSIONS

Crime, as we have seen, is a rich and complex object of study. Its incidence depends not just on police and prisons, but on the air we breathe (or breathed when we were children), what our mothers drank, what our windows are made of, how we enter our cars, how we pay our bills, whether the home football team won, and so on. The topic is rich because crime covers a wide array of activities—from poisoning a former lover's goldfish to robbing a bank to joyriding in a borrowed car to assassinating a rival drug dealer—that are joined together only by being proscribed. But it is also rich because of the complexity of modern urban life. General equilibrium, rather than incentives, may turn out to be the big idea that economics brings to the study of crime.

Because crime is tightly integrated with the rest of life, crime changes when technology changes. We emphasized a positive role for technology when we discussed the last two decades' drops in victimization. But if we leave aside the restriction to index crimes, it is not absurd to argue that we are currently experiencing the worst crime wave ever. For 2012, the NCVS found that 26.5 million index crime victimizations occurred in the United States, but in just a few weeks from November 27 to mid December 2013, between 70 million and 110 million people had valuable financial information stolen from them at a single source (Harris and Perlroth, 2014). Phishing e-mails and online scams can target millions of potential victims at a stroke, while cyberstalking and cyberbullying can be as devastating to a victim as larceny or assault.

What the future brings, we have no idea. Civilian drones may help some police departments, but they can also transport drugs, enter houses and steal jewelry, and spy on police and security guards. The first civilian murder by drone may very well occur before this volume is published. Autonomous cars similarly open up exciting new opportunities for carjacking. Technology is a horse race between those who want to commit crimes and those who do not want them to, and we are not placing any bets.

The character of crime will change, and with it will change the nature of the questions that economists are called upon to answer. We cannot predict these developments with much confidence, but of one thing we can be sure: the conflict between Lombroso and Beccaria, between incapacitation and deterrence, between prediction and detection will remain central to our understanding of crime and punishment.

ACKNOWLEDGMENTS

Written for the *Handbook of Urban and Regional Economics*, edited by Gilles Duranton, Vernon Henderson and William Strange. In addition to the editors, we thank Douglas Almond, Jeff Fagan, Ally Fedorov, Yiming He, John MacDonald, Robert McMillan, and Steven Raphael for comments on earlier versions, and Suxin Shen for research assistance.

REFERENCES

- Abadie, A., Dermisi, S., 2008. Is terrorism eroding agglomeration economies in central business districts? Lessons from the office real estate market in downtown Chicago. *J. Urban Econ.* 64, 451–463.
- Aizer, A., Doyle Jr., J.J., 2013. Juvenile incarceration, human capital and future crime: evidence from randomly-assigned judges. National Bureau of Economic Research.
- Aizer, A., Stroud, L., 2011. Education, knowledge, and the evolution of disparities in health. Brown University, Working paper.
- Akinbami, O.J., Liu, X., Pastor, P.N., Reuben, C.A., 2011. Attention Deficit Hyperactivity Disorder Among Children Aged 5–17 Years in the United States, 1998–2009. US Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Health Statistics.
- Alexander, M., 2010. *The New Jim Crow: Mass Incarceration in the Age of Colorblindness*. The New Press.
- Altonji, J.G., Doraszelski, U., 2005. The role of permanent income and demographics in black/white differences in wealth. *J. Hum. Res.* 40, 1–30.

- Ananat, E.O., Gruber, J., Levine, P., Staiger, D., 2006. Abortion and selection. NBER Working paper 12150.
- Ananat, E.O., Gruber, J., Levine, P., Staiger, D., 2009. Abortion and selection. *Rev. Econ. Stat.* 91, 124–136.
- Anderson, D.A., 1999. The aggregate burden of crime. *J. Law Econ.* 42, 611–642.
- Antecol, H., Bedard, K., 2007. Does single parenthood increase the probability of teenage promiscuity, substance use, and crime? *J. Popul. Econ.* 20, 55–71.
- Anwar, S., Fang, H., 2006. An alternative test of racial prejudice in motor vehicle searches: theory and evidence. *Am. Econ. Rev.* 96, 127–151.
- Arbel, Y., Ben-Shahar, D., Gabriel, S., Tobol, Y., 2010. The local cost of terror: effects of the second Palestinian Intifada on Jerusalem house prices. *Reg. Sci. Urban Econ.* 40, 415–426.
- Ayres, I., 2001. *Pervasive Prejudice? Unconventional Evidence of Race and Gender Discrimination*. University of Chicago Press, Chicago, IL.
- Ayres, I., Levitt, S.D., 1998. Measuring positive externalities from unobservable victim precaution: an empirical analysis of Lojack. *Q. J. Econ.* 113, 43–77.
- Baliga, S., Sjöström, T., 2004. Arms races and negotiations. *Rev. Econ. Stud.* 71, 351–369.
- Barbarino, A., Mastrobuoni, G., 2014. The incapacitation effect of incarceration: evidence from several Italian collective pardons. *Am. Econ. J. Appl. Econ.* 6, 1–37.
- Bayer, P., Hjalmarsson, R., Pozen, D., 2009. Building criminal capital behind bars: peer effects in juvenile corrections. *Q. J. Econ.* 124.
- Beccaria, C., 1764. *On Crimes and Punishments (Dei Delitti e Della Pene)*. James Donaldson, Edinburgh, English translation, 1788.
- Becker, G., 1968. Crime and punishment: an economic approach. *J. Pol. Econ.* 76, 169–217.
- Becker, G.S., 1957. *The Economics of Discrimination*. University of Chicago Press, Chicago, IL.
- Berk, R.A., Rauma, D., 1983. Capitalizing on nonrandom assignment to treatments: a regression-discontinuity evaluation of a crime-control program. *J. Am. Stat. Assoc.* 78, 21–27.
- Bjerk, D., 2007. Racial profiling, statistical discrimination, and the effect of a colorblind policy on the crime rate. *J. Publ. Econ. Theor.* 9.
- Blau, F.D., Graham, J.W., 1990. Black-white differences in wealth and asset composition. *Q. J. Econ.* 105, 321–339.
- Blomberg, S.B., Sheppard, S., 2007. The impacts of terrorism on urban form. *Brook. Wharton Pap. Urban Aff.* 257–290, 295–296.
- Blow, C., 2013. The whole system failed Trayvon Martin. *N. Y. Times*. (July 15).
- Blumstein, A., Wallman, J. (Eds.), 2000. *The Crime Drop in America*. Cambridge University Press, New York, NY.
- Blumstein, A., Wallman, J. (Eds.), 2006a. *The Crime Drop in America*. Revised Edition Cambridge University Press, New York, NY.
- Blumstein, A., Wallman, J., 2006b. The crime drop and beyond. *Ann. Rev. Law Soc. Sci.* 125–146.
- Braga, A., 2008. Crime Prevention Research Review No. 2: Police Enforcement Strategies to Prevent Crime in Hot Spot Areas. U.S. Department of Justice, Office of Community Policing Strategies, Washington, D.C.
- Buonanno, P., Leonida, L., 2006. Education and crime: evidence from Italian regions. *Appl. Econ. Lett.* 13, 709–713.
- Buonanno, P., Raphael, S., 2013. Incarceration and incapacitation: evidence from the 2006 Italian collective pardon. *Am. Econ. Rev.* 103, 2437–2465.
- Bureau of Justice Statistics., Sourcebook of criminal justice statistics, various years.
- Bushway, S., Paternoster, R., 2009. The impact of prison on crime, in: *Do Prisons Make Us Safer? The Benefits and Costs of the Prison Boom*.
- Bushway, S.D., Stoll, M.A., Weiman, D. (Eds.), 2007. *Barriers to Reentry? The Labor Market for Released Prisoners in Post-industrial America*. Russell Sage Foundation, New York, NY.
- Card, D., Mas, A., Rothstein, J., 2008. Tipping and the dynamics of segregation. *Q. J. Econ.* 123, 177–218.
- Carneiro, P., Ginja, R., 2008. Preventing behavior problems in childhood and adolescence. University of Pennsylvania Working paper.

- Chalfin, A., McCrary, J., 2013. The effect of police on crime: new evidence from U.S. cities, 1960–2010. National Bureau of Economic Research Working paper 18815.
- Charles, K.K., Hurst, E., Roussanov, N., 2009. Conspicuous consumption and race. *Q. J. Econ.* 124, 425–467.
- Chen, B.X., 2014. Smartphones embracing theft defense. *N. Y. Times*. (June 20).
- Chen, M.K., Shapiro, J.M., 2007. Do harsher prison conditions reduce recidivism? A discontinuity-based approach. *Am. Law Econ. Rev.* 9, 1–29.
- Cheng, C., Hoekstra, M., 2012. Does strengthening self-defense law deter crime or escalate violence? Evidence from castle doctrine. National Bureau of Economic Research Working paper.
- Choe, J.Y., Teplin, L.A., Abram, K.M., 2008. Perpetration of violence, violent victimization, and severe mental illness: balancing public health concerns. *Psychiatr. Serv.* 59, 153–164.
- Clear, T.R., 2007. *Imprisoning Communities: How Mass Incarceration Makes Disadvantaged Neighborhoods Worse*. Oxford University Press.
- Coase, R.H., 1960. The problem of social cost. *J. Law Econ.* 3, 1–44.
- Cook, P., Ludwig, J., 2006. Aiming for evidence-based gun policy. *J. Pol. Anal. Manag.* 25, 691–735.
- Cook, P.J., 1983. Encyclopedia of crime and justice. In: Kadish, S.H. (Ed.), *Encyclopedia of Crime and Justice*. Free Press, New York, NY.
- Cook, P.J., 1986. The demand and supply of criminal opportunities. *Crime Justice* 7, 1–27.
- Cook, P.J., Ludwig, J., 2010. Economical crime control. National Bureau of Economic Research Working paper.
- Cook, P.J., MacDonald, J., 2011. The role of private action in controlling crime. In: Cook, P.J., Ludwig, J., McCrary, J. (Eds.), *Controlling Crime: Strategies and Tradeoffs*. University of Chicago Press.
- Corman, H., Dave, D., Reichman, N., 2013. Effects of welfare reform on women's crime. National Bureau of Economic Research Working paper 18887.
- Coviello, D., Persico, N., 2013. An economic analysis of black-white disparities in NYPD's stop and frisk program. NBER Working papers.
- Cuellar, A.E., Markowitz, S., 2007. Medicaid policy changes in mental health care and their effect on mental health outcomes. *Health Econ. Pol. Law* 2.
- Cullen, J., Jacob, B., Levitt, S., 2006. The effect of school choice on participants: evidence from randomized lotteries. *Econometrica* 74, 1191–1230.
- Curtis, M., Garlington, S., Schottenfield, L.S., 2013. Alcohol, drug, and criminal history restrictions in public housing. *Cityscape J. Pol. Dev. Res.* 15, 37–52.
- Cutler, D.M., Glaeser, E.L., Vigdor, J.L., 1999. The rise and decline of the American Ghetto. *J. Pol. Econ.* 107, 455–506.
- Dell, M., 2012. *Trafficking networks and the Mexican drug war*. Harvard University, Working paper.
- Deming, D., 2009. Early childhood intervention and life-cycle skill development. *Am. Econ. J. Appl. Econ.* 1, 111–134.
- Deming, D.J., 2011. Better schools, less crime? *Q. J. Econ.* 126, 2063–2115.
- Dharmapala, D., Ross, S.L., 2004. Racial bias in motor vehicle searches: additional theory and evidence. *Contr. Econ. Anal. Pol.* 3.
- Di Tella, R., Schargrodsky, E., 2004. Do police reduce crime? Estimates using the allocation of police forces after a terrorist attack. *Am. Econ. Rev.* 94, 115–133.
- Di Tella, R., Schargrodsky, E., 2009. Criminal recidivism after prison and electronic monitoring. National Bureau of Economic Research Working paper.
- Donohue, J., 2009. Assessing the relative benefits of incarceration: the overall change over the previous decades and the benefits on the margin. In: *Do Prisons Make Us Safer? The Benefits and Costs of the Prison Boom*.
- Donohue, J., Levitt, S., 2001. The impact of legalized abortion on crime. *Q. J. Econ.* 116, 379–420.
- Donohue, J., Levitt, S.D., 2004. Further evidence that legalized abortion lowered crime: a reply to Joyce. *J. Hum. Res.* 39.
- Donohue, J.J., Levitt, S.D., 2008. Measurement error, legalized abortion, and the decline in crime: a response to Foote and Goetz. *Q. J. Econ.* 123, 425–440.

- Draca, M., Machin, S., Witt, R., 2011. Panic on the streets of London: police, crime, and the July 2005 terror attacks. *Am. Econ. Rev.* 101, 2157–2181.
- Drago, F., Galbiati, R., 2012. Indirect effects of a policy altering criminal behaviour: evidence from the Italian prison experiment. *Am. Econ. J. Appl. Econ.* 4, 199–218.
- Drago, F., Galbiati, R., Vertova, P., 2009. The deterrent effects of prison: evidence from a natural experiment. *J. Pol. Econ.* 117.
- Durlauf, S.N., Nagin, D.S., 2011. The deterrent effect of imprisonment. In: Cook, P.J., Ludwig, J., McCrary, J. (Eds.), *Controlling Crime: Strategies and Tradeoffs*. University of Chicago Press, Chicago, IL, pp. 43–94.
- Easterlin, R., 1973. Relative economic status and the American fertility swing. In: Sheldon, E. (Ed.), *Family Economic Behavior: Problems and Prospects*. Lippincott, Philadelphia, PA.
- Eck, J.E., Maguire, E., 2000. Have changes in policing reduced violent crime? An assessment of the evidence. In: Blumstein, Wallman (Ed.), *The Crime Drop in America*. Cambridge University Press, New York, NY.
- Ehrlich, I., 1973. Participation in illegitimate activities: a theoretical and empirical investigation. *J. Pol. Econ.* 81, 521–565.
- Ellison, G., Glaeser, E.L., 1997. Geographic concentration in U.S. manufacturing industries: a dartboard approach. *J. Pol. Econ.* 105, 889–927.
- European Sourcebook of Crime and Criminal Justice Statistics, 2010. Universite de Lausanne, Institut de criminology and de droit penal.
- Evans, W.N., Owens, E.G., 2007. COPS and crime. *J. Publ. Econ.* 91, 181–201.
- Fagan, J., 2010. Expert report submitted to United States district court southern district of New York in *floyd v. city of new york*.
- Fajnzylber, P., Lederman, D., Loayza, N., 2002. What causes violent crime? *Eur. Econ. Rev.* 46, 1323–1357.
- Farley, R., Frey, W.H., 1994. Changes in the segregation of whites from blacks during the 1980s: small steps toward a more integrated society. *Am. Sociol. Rev.* 59, 23–45.
- Fleisher, M.S., Decker, S.H., 2001. An overview of the challenge of prison gangs. *Corr. Manag. Q.* 5.
- Frank, R.G., McGuire, T.G., 2011. Mental health treatment and criminal justice outcomes. In: Cook, P., Ludwig, J., McCrary, J. (Eds.), *Controlling Crime: Strategies and Payoffs*. University of Chicago Press, Chicago, IL.
- Freeman, R.B., 1996. Why do so many young American men commit crimes and what might we do about it? *J. Econ. Perspect.* 10, 25–42.
- Freeman, R.B., 2001. Does the booming economy help explain the fall in crime?, in: *Perspectives in Crime and Justice: 1999–2000 Lecture Series*, NCJ 184245. National Institute of Justice, Washington, D.C., pp. 23–43.
- Fryer, R.G., Heaton, P.S., Levitt, S.D., Murphy, K.M., 2013. Measuring crack cocaine and its impact. *Econ. Inq.* 51, 1651–1681.
- Furman Center for Real Estate and Urban Policy, 2007. The benefits of business improvement districts: evidence from New York City. Policy Brief, New York University.
- Galliani, S., Rossi, M.A., Scharfrodsky, E., 2011. Conscription and crime: evidence from the Argentine draft lottery. *Am. Econ. J. Appl. Econ.* 3, 119–136.
- Garces, E., Thomas, D., Currie, J., 2002. Longer-term effects of head start. *Am. Econ. Rev.* 92, 999–1012.
- Gautier, P.A., Siegmund, A., Vuuren, A.V., 2009. Terrorism and attitudes towards minorities: the effect of the Theo van Gogh murder on house prices in Amsterdam. *J. Urban Econ.* 65, 113–126.
- Gelman, A., Fagan, J., Kiss, A., 2007. An analysis of the New York city police department's "stop-and-frisk" policy in the context of claims of racial bias. *J. Am. Stat. Assoc.* 102, 813–823.
- Glaeser, E.L., 2007. The impacts of terrorism on urban form [comment]. *Brook. Wharton Pap. Urban Aff.* 291–294.
- Glaeser, E.L., Sacerdote, B., 1999. Why is there more crime in cities? *J. Pol. Econ.* 107, S225–S258.
- Glaeser, E.L., Sacerdote, B., Scheinkman, J.A., 1996. Crime and social interactions. *Q. J. Econ.* 111, 507–548.

- Glaeser, E.L., Shapiro, J.M., 2002. Cities and warfare: the impact of terrorism on urban form. *J. Urban Econ.* 51, 205–224.
- Glaze, L.E., 2010. Correctional populations in the United States 2009. US Bur. Just. Stat. NCJ 231681.
- Gould, E.D., Weinberg, B.A., Mustard, D.B., 2002. Crime rates and local labor market opportunities in the United States: 1979–1997. *Rev. Econ. Stat.* 84, 45–61.
- Grogger, J., 1998. Market wages and youth crime. *J. Lab. Econ.* 16, 756–791.
- Grosjean, P., 2014. A history of violence: the culture of honor and homicide in the U.S. South. *J. Eur. Econ. Assoc.* forthcoming.
- Harbaugh, W.T., Mocan, N.H., Visser, M.S., 2013. Theft and deterrence. *J. Lab. Res.* 34, 89–407.
- Harcourt, B.E., 2006. *Against Prediction: Profiling, Policing, and Punishing in an Actuarial Age.* University of Chicago Press.
- Harris, E.A., 2014. For target, the breach numbers grow. *N. Y. Times* January 11.
- Helland, E., Tabarrok, A., 2007. Does three strikes deter? A nonparametric estimation. *J. Hum. Res.* 42, 309–330.
- Heller, S., Pollack, H.A., Ander, R., Ludwig, J., 2013. Preventing youth violence and dropout: a randomized field experiment. National Bureau of Economic Research Working paper 19014.
- Helsley, R.W., Strange, W.C., 1999. Gated communities and the economic geography of crime. *J. Urban Econ.* 46, 80–105.
- Helsley, R.W., Strange, W.C., 2005. Mixed markets and crime. *J. Publ. Econ.* 89, 1251–1275.
- Hill, P.L., Roberts, B.W., Grogger, J.T., Guryan, J., Sixkiller, K., 2011. Decreasing delinquency, criminal behavior, and recidivism by intervening on psychological factors other than cognitive ability: a review of the intervention literature. In: Cook, P., Ludwig, J., McCrary, J. (Eds.), *Controlling Crime: Strategies and Payoffs.* University of Chicago Press, Chicago, IL, National Bureau of Economic Research Conference Report.
- Hirschman, A.O., 1970. *Exit, Voice, and Loyalty: Responses to Decline in Firms, Organizations, and States.* Harvard University Press.
- Holzer, H.J., 2009. Collateral costs: effects of incarceration on employment and earnings among young workers. In: Raphael, S., Stoll, M.A. (Eds.), *Do Prisons Make Us Safer? The Benefits and Costs of the Prison Boom.* Russell Sage, New York, NY, pp. 239–268.
- Hotelling, H., 1929. Stability in competition. *Econ. J.* 39, 41–57.
- Imrohoroglu, A., Merlo, A., Rupert, P., 2004. What accounts for the decline in crime? *Int. Econ. Rev.* 45, 707–729.
- International Centre for Prison Studies, 2011. *World prison population list (ninth edition).*
- Jacob, B.A., Lefgren, L., 2003. Are idle hands the devil's workshop? Incapacitation, concentration and juvenile crime. National Bureau of Economic Research.
- Jacobs, J., 1961. *The Death and Life of Great American Cities.* Random House LLC.
- Johnson, R., 2009. Ever-increasing levels of parental incarceration and the consequences for children. In: *Do Prisons Make Us Safer? The Benefits and Costs of the Prison Boom.* Russell Sage Foundation, New York, NY, pp. 177–206.
- Johnson, R.C., Raphael, S., 2012. How much crime reduction does the marginal prisoner buy? *J. Law Econ.* 55, 275–310.
- Johnson, R.S., Kantor, S., Fishback, P.V., 2007. Striking at the roots of crime: the impact of social welfare spending on crime during the great depression. NBER Working paper 12825.
- Jones, K., Smith, D., 1973. Recognition of the fetal alcohol syndrome in early infancy. *Lancet* 2, 999–1001.
- Joyce, T.J., 2009. Abortion and crime: a review. National Bureau of Economic Research Working paper 15098.
- Katz, L., Levitt, S.D., Shustorovich, E., 2003. Prison conditions, capital punishment, and deterrence. *Am. Law Econ. Rev.* 5 (2), 318–343.
- Kaufman, J., 2013. The building has 1,000 eyes: residences large and small, doorman or no, are stepping up their reliance on surveillance cameras. *N. Y. Times* October 6.
- Kendall, T.D., Tamura, R., 2010. Unmarried fertility, crime and social stigma. *J. Law Econ.* 53, 185–221.

- Kennedy, D., Braga, A., Piehl, A., Waring, E., 2001. Reducing Gun Violence: The Boston Gun Project's Operation Ceasefire. National Institute of Justice, Washington, D.C.
- Kessler, R.C., Adler, L., Barkley, R., Biederman, J., Conners, K., Demler, O., Faraone, S.V., Greenhill, L. L., Howes, M.J., Secnik, K., Spencer, T., Ustun, B., Walters, E.E., Zavilavsky, A., 2006. The prevalence and correlates of adult ADHD in the United States: results from the National Comorbidity Survey Replication. *Am. J. Psychiat.* 163, 716–723.
- Kleiman, M., 2009. When Brute Force Fails: How to Have Less Crime and Less Punishment. Princeton University Press.
- Klick, J., Tabarrok, A., 2005. Using terror alert levels to estimate the effect of police on crime. *J. Law Econ.* 48, 267–279.
- Kling, J.R., 2006. Incarceration length, employment, and earnings. *Am. Econ. Rev.* 96, 863–876, National Bureau of Economic Research Working paper 12003.
- Knowles, J., Persico, N., Todd, P., 2001. Racial bias in motor vehicle searches: theory and evidence. *J. Pol. Econ.* 109, 203–229.
- Kocieniewski, D., 2007. A little girl shot, and a crowd that didn't see. *N. Y. Times* July 9.
- Kutateladze, B., Lynn, V., Liang, E., 2012. Do race and ethnicity matter in prosecution? A review of empirical studies. Vera Institute of Justice, New York.
- Kuziemko, I., Levitt, S.D., 2004. An empirical analysis of imprisoning drug offenders. *J. Publ. Econ.* 88.
- Lee, D., McCrary, J., 2009. The deterrent effect of prison. Princeton University Department of Economics Working paper.
- Lee, S., 2008. The impact of burglar alarm systems on residential burglaries. Ph.D. thesis, In: Rutgers School of Criminal Justice.
- Lerman, A.E., 2009. The people prisons make: effects of incarceration on criminal psychology. In: *Do Prisons Make Us Safer?: The Benefits and Costs of the Prison Boom*. Russell Sage Foundation, pp. 151–176.
- Levine, P.B., Staiger, D., Kane, J., Zimmerman, D., 1999. Roe v. Wade and American fertility. *Am. J. Publ. Health* 89, 199–203.
- Levitt, S.D., 1996. The effect of prison population size on crime rates: evidence from prison overcrowding litigation. *Q. J. Econ.* 111.
- Levitt, S.D., 1997. Using electoral cycles in police hiring to estimate the effect of police on crime. *Am. Econ. Rev.* 270–290.
- Levitt, S.D., 1998. Why do increased arrest rates appear to reduce crime? Deterrence, incapacitation, or measurement error. *Econ. Inq.* 36 (3), 353–372.
- Levitt, S.D., 1998. Juvenile crime and punishment. *J. Pol. Econ.* 106 (6), 1156–1185.
- Levitt, S.D., 1999. The limited role of changing age structure in explaining aggregate crime rates. *Criminology* 37, 581–597.
- Levitt, S.D., 2002. Using election cycles in police hiring to estimate the effect of police on crime: reply. *Am. Econ. Rev.* 87 (3), 270–290.
- Levitt, S.D., 2004. Understanding why crime fell in the 1990s: four factors that explain and six that do not. *J. Econ. Perspect.* 18, 163–190.
- Liedka, R.V., Piehl, A.M., Useem, B., 2006. The crime control effect of incarceration: does scale matter? *Crim. Publ. Pol.* 5, 245–276.
- Lin, M.J., 2009. More police, less crime: evidence from US state data. *Int. Rev. Law Econ.* 29 (2), 73–80.
- Lochner, L., 2007. Individual perceptions of the criminal justice system. *Am. Econ. Rev.* 97, 444–460.
- Lochner, L., 2011. Education and crime. In: Cook, P., Ludwig, J., McCrary, J. (Eds.), *Controlling Crime: Strategies and Payoffs*. University of Chicago Press, Chicago, IL, National Bureau of Economic Research Conference Report.
- Lochner, L., Moretti, E., 2004. The effect of education on crime: evidence from prison inmates, arrests, and self-reports. *Am. Econ. Rev.* 94.
- Loeffler, C.E., 2013. Does imprisonment alter the life course? Evidence on crime and employment from a natural experiment. *Criminology* 51, 137–166.
- Loury, G.C., 2002. *The Anatomy of Racial Inequality*. Harvard University Press.
- Loury, G.C., 2008. *Race, Incarceration, and American Values*. MIT Press.

- Loury, G.C., 2009. A nation of jailers. *Cato Unbound*.
- Lu, R., Lin, W., Lee, J., Ko, H., Shih, J., 2003. Neither antisocial personality disorder nor antisocial alcoholism is associated with MAO-A gene in Han Chinese males. *Alcoh. Clin. Exper. Res.* 27 (6), 889–893.
- Luallen, J., 2006. School's out... forever: a study of juvenile crime, at-risk youths and teacher strikes. *J. Urban Econ.* 59, 75–103.
- Machin, S., Marie, O., Vujic, S., 2011. The crime reducing effect of education. *Econ. J.* 121, 463–484.
- Marcotte, D., Markowitz, S., 2011. A cure for crime? Psychopharmaceuticals and crime trends. *J. Pol. Anal. Manag.* 30, 29–56.
- Massey, D.S., Denton, N.A., 1987. Trends in the residential segregation of blacks, hispanics, and asians: 1970–1980. *Am. Sociol. Rev.* 52, 802–825.
- McClellan, C.B., Tekin, E., 2012. Stand your ground laws, homicides, and injuries.
- McCrary, J., 2002. Using electoral cycles in police hiring to estimate the effect of police on crime: comment. *Am. Econ. Rev.* 1236–1243.
- McFadden, R.D., Saulny, S., 2002. A crime revisited: the decision; 13 years later, official reversal in Jogger Attack. *N.Y. Times*.
- Mills, E.S., 2002. Terrorism and US real estate. *J. Urban Econ.* 51, 198–204.
- Minton, T.D., 2012. Jail inmates at midyear 2011: statistical tables. NCJ237961, Bur. Just. Stat. Accessed at, <http://www.bjs.gov/content/pub/pdf/jim11st.pdf>.
- Miron, J., 1999. Violence and the U.S. prohibitions of drugs and alcohol. *Am. Law Econ. Rev.* 1, 78–114.
- Moffitt, T., Ross, S., 2011. Enhancing children's self-control to reduce crime and promote the health and welfare of the population. In: Cook, P., Ludwig, J., McCrary, J. (Eds.), *Controlling Crime: Strategies and Payoffs*. University of Chicago Press, Chicago IL, National Bureau of Economic Research Conference Report.
- Myrdal, G., 1944. *American Dilemma: The Negro Problem and Modern Democracy*. Harper and Brothers.
- Nagin, D., 1998. Criminal deterrence: research at the outset of the 21st century. *Crime Justice* 23, 1–42.
- Nagin, D.S., Cullen, F.T., Jonson, C.L., 2009. Imprisonment and reoffending. *Crime Justice* 38, 115–200.
- Neugebauer, R., Hoek, H., Susser, E., 1999. Prenatal exposure to wartime famine and development of antisocial personality disorder in early adulthood. *J. Am. Med. Assoc.* 4, 479–481.
- Newport, F., 2010. US drinking rate edges up slightly to 25-year high. Gallup Poll, accessed at www.gallup.com/poll/141656/drinking-rate-edges-slightly-year-high.aspx.
- O'Brien, R., 1989. Relative cohort size and age-specific crime rates: an age-period-relative-cohort-size model. *Criminology* 27, 57–77.
- O'Flaherty, B., 1998. Why repeated criminal opportunities matter: a dynamic stochastic model of criminal decision-making. *J. Law Econ. Organ.* 14 (2), 232–255.
- O'Flaherty, B., Sethi, R., 2007. Crime and segregation. *J. Econ. Behav. Organ.* 64, 391–405.
- O'Flaherty, B., Sethi, R., 2008. Racial stereotypes and robbery. *J. Econ. Behav. Organ.* 68, 511–524.
- O'Flaherty, B., Sethi, R., 2009a. Why have robberies become less frequent but more violent? *J. Law Econ. Organ.* 25, 518–534.
- O'Flaherty, B., Sethi, R., 2009b. Public outrage and criminal justice: lessons from the Jessica Lal case. In: Dutta, B., Ray, T., Somanathan, E. (Eds.), *New and Enduring Themes in Development Economics*. World Scientific.
- O'Flaherty, B., Sethi, R., 2010a. The racial geography of street vice. *J. Urban Econ.* 67, 270–286.
- O'Flaherty, B., Sethi, R., 2010b. Witness intimidation. *J. Legal Stud.* 39, 399–432.
- O'Flaherty, B., Sethi, R., 2010c. Peaceable kingdoms and war zones: preemption, ballistics and murder in Newark. In: Tella, R.D., Edwards, S., Schargrodsky, E. (Eds.), *The Economics of Crime: Lessons for and from Latin America*. University of Chicago Press.
- O'Flaherty, B., Sethi, R., 2010d. Homicide in black and white. *J. Urban Econ.* 68, 215–230.
- Oliver, M.L., Shapiro, T.M., 1995. *Black Wealth White Wealth: A New Perspective on Racial Inequality*. Routledge.
- Owens, E.G., 2009. More time, less crime? Estimating the incapacitative effect of sentence enhancements. *J. Law Econ.* 52, 551–579.
- Pager, D., Western, B., Sugie, N., 2009. Sequencing disadvantage: barriers to employment facing young black and white men with criminal records. *Ann. Am. Acad. Pol. Soc. Sci.* 623, 195–213.

- Papachristou, A., Meares, T., Fagan, J., 2007. Attention, felons: evaluating project safe neighborhood in Chicago. *J. Emp. Leg. Stud.* 4, 223–272.
- Patterson, O., 1982. *Slavery and Social Death: A Comparative Study*. Harvard University Press.
- Perreault, S., 2013. Police reported crime statistics in Canada 2012. *Juristat*, Statistics Canada.
- Pirog, M.A., Johnson, C.L., 2008. Electronic funds and benefits transfers, e-government, and the Winter Commission. *Publ. Admin. Rev.* 68, S103–S114.
- Ponseti, J., Granert, O., van Eimeren, T., Jansen, O., Wolff, S., Beier, K., Deuschl, G., Bosinski, H., Siebner, H., 2014. Human face processing is tuned to sexual age preferences. *Biol. Lett.* 10.
- Poutvaara, P., Priks, M., 2009. The effect of police intelligence on group violence: evidence from reassignments in Sweden. *J. Publ. Econ.* 93, 403–411.
- Pratt, T.C., Cullen, F.T., 2005. Assessing macro-level predictors and theories of crime: a meta-analysis. In: Tonry, M. (Ed.), *Crime and Justice: A Review of Research*. University of Chicago Press, Chicago, IL, volume 32.
- Raine, A., 2013. *The Anatomy of Violence: The Biological Roots of Crime*. Pantheon, New York, NY.
- Raphael, S., Ludwig, J., 2003. Prison sentence enhancement: the case of Project Exile. In: Ludwig, J., Cook, P. (Eds.), *Evaluating Gun Policy: Effects on Crime and Violence*. Brookings Institution Press, Washington, D.C, pp. 251–286.
- Raphael, S., Stoll, M., 2013. Why Are So Many Americans in Prison? Russell Sage, New York, NY.
- Raphael, S., Stoll, M.A., 2009a. Introduction. In: Raphael, S., Stoll, M.A. (Eds.), *Do Prisons Make Us Safer? The Benefits and Costs of the Prison Boom*. Russell Sage, New York, NY, pp. 1–26.
- Raphael, S., Stoll, M.A., 2009b. Why are so many Americans in prison? In: Raphael, S., Stoll, M.A. (Eds.), *Do Prisons Make Us Safer? The Benefits and Costs of the Prison Boom*. Russell Sage, New York, NY, pp. 27–72.
- Raphael, S., Winter-Ebmer, R., 2001. Identifying the effect of unemployment on crime. *J. Law Econ.* 44, 259–283.
- Reiss Jr., A.J., Roth, J.A. (Eds.), 1993. *Understanding and Preventing Violence*. National Academy Press, Washington, D.C.
- Reyes, J.W., 2007. Environmental policy as social policy? The impact of childhood lead exposure on crime. *BE J. Econ. Anal. Pol.* 7.
- Reyes, J.W., 2012. Lead exposure and behavior: effects on antisocial and risky behavior among children and adolescents. Amherst College, Working paper.
- Reynolds, A., Temple, J., Robertson, D., Mann, E., 2001. Long-term effects of an early childhood intervention on educational achievement and juvenile arrest: a 15-year follow-up of low-income children in public schools. *J. Am. Med. Assoc.* 285, 2239–2246.
- Rice, D., 2014. USA saw fewest lightning deaths on record in 2013. *USA Today*, January 11.
- Riolo, S.A., Nguyen, T.A., Greden, J.F., King, C.A., 2005. Prevalence of depression by race/ethnicity: findings from the National Health and Nutrition Examination Survey III. *Am. J. Publ. Health* 95 (6), 998–1000.
- Sabol, S.Z., Hu, S., Hamer, D., 1998. A functional polymorphism in the monoamine oxidase A gene promoter. *Human Genet.* 103, 273–279.
- Salop, S., 1979. Monopolistic competition with outside goods. *Bell J. Econ.* 10, 145–156.
- Sampson, R.J., Cohen, J., 1988. Deterrent effects of police on crime: a replication and theoretical extension. *Law Soc. Rev.* 22.
- Sampson, R.J., Lauritsen, J., 1997. Racial and ethnic disparities in crime and criminal justice in the United States. *Crime Justice* 21, 311–374.
- Sampson, R.J., Raudenbush, S.W., Earls, F., 1997. Neighborhoods and violent crime: a multilevel study of collective efficacy. *Science*. 918.
- Sanga, S., 2009. Reconsidering racial bias in motor vehicle searches: theory and evidence. *J. Pol. Econ.* 117, 1155–1159.
- Santora, M., Schwartz, M., 2013. An epidemic of carjackings afflicts Newark. *N. Y. Times* (December 28).
- Saperstein, A., Panner, A., 2010. The race of a criminal record: how incarceration colors racial perception. *Soc. Probl.* 57 (1), 92–113.
- Schelling, T.C., 1971. Dynamic models of segregation. *J. Math. Sociol.* 1, 143–186.
- Schrag, J., Scotchmer, S., 1997. The self-reinforcing nature of crime. *Int. Rev. Law Econ.* 17, 325–335.

- Sethi, R., Somanathan, R., 2004. Inequality and segregation. *J. Pol. Econ.* 112, 1296–1321.
- Sethi, R., Somanathan, R., 2009. Racial inequality and segregation measures: some evidence from the 2000 census. *Rev. Black Pol. Econ.* 36, 79–91.
- Sherman, L.W., Gartin, P.R., Buerger, M.E., 1989. Hot spots of predatory crime: routine activities and the criminology of place. *Criminology* 27, 27–56.
- Skarbek, D., 2012. Prison gangs, norms, and organizations. *J. Econ. Behav. Organ.* 82, 96–109.
- Skogan, W., Frydl, K. (Eds.), 2004. *Fairness and Effectiveness in Policing: The Evidence* (Committee to Review Research on Police Policy and Practices). The National Academies Press, Washington, D.C.
- Spelman, W., 2005. Jobs or jails? The crime drop in Texas. *J. Pol. Anal. Manag.* 24, 133–165.
- Statistical Abstract, United States bureau of the census, various years.
- Steffensmeier, D., Striefel, C., Harer, M., 1987. Relative cohort size and youth crimes in the United States, 1953–1984. *Am. Sociol. Rev.* 52, 702–710.
- Stegman, M.A., Lobenhofer, J.S., Quinterno, J., 2003. The state of electronic benefits transfer (EBT). Center for Community Capital, University of North Carolina at Chapel Hill, Working paper.
- The Sentencing Project, 2011. Trends in US corrections.
- Tillman, R., 1987. The size of the ‘criminal population’: the prevalence and incidence of adult arrest. *Criminology* 25, 561–579.
- United States General Accounting Office, 2000. Racial profiling: limited data available on most motorist stops.
- U.S. Bureau of Labor Statistics, 2013b. Occupational employment statistics.
- U.S. Department of Health and Human Services, 1991. National household survey of drug abuse, highlights.
- U.S. District Court, Southern District of New York, 2013. *United States of America v. Andrew Michael Jones aka “Inigo”, Gary Davis, aka “Libertas”, Peter Philip Nash, aka “samesamebutdifferent”, aka “Batman73”, aka “Symmetry”, aka “Anonymousasshit”*, S1 13 Cr. 950.
- U.S. Social Security Administration, 2000. History of the social security administration 1993–2000. Accessed at www.ssa.gov/history/ssa/ssa2000chapter5.html.
- Valla, J.M., Ceci, S.J., Williams, W., 2011. The accuracy of inferences about criminality based on facial appearance. *J. Soc. Evolut. Cult. Psychol.* 5 (1), 66–91.
- van Ours, J.C., Vollaard, B., 2013. The engine immobilizer: a non-starter for car thieves. CESifo Working paper: Public Choice.
- Verdier, T., Zenou, Y., 2004. Racial beliefs, location, and the causes of crime. *Int. Econ. Rev.* 45, 731–760.
- Vollaard, B., 2013. Preventing crime through selective incapacitation. *Econ. J.* 123, 262–284.
- Vollaard, B., van Ours, J., 2011. Does regulation of built-in security reduce crime? Evidence from a natural experiment. *Econ. J.* 121, 485–504.
- Washington, B.T., 1915. My view of segregation laws. New Republ.
- Welsh, B.C., Farrington, D.P., 2009. Effects of closed-circuit television on crime. *Ann. Am. Acad. Pol. Soc. Sci.* 587, 110–135.
- Widom, C., Brzustowicz, L., 2006. MAOA and the icycle of violence: childhood abuse and neglect, MAOA genotype, and risk for violent and antisocial behavior. *Biol. Psychiat.* 60, 684–689.
- Wildasin, D.E., 2002. Local public finance in the aftermath of September 11. *J. Urban Econ.* 51, 225–237.
- Wildeman, C., 2010. Paternal incarceration and children’s physically aggressive behaviors: evidence from the fragile families and child wellbeing study. *Soc. Forc.* 89, 285–309.
- Wilson, J.Q., Kelling, G., 1982. Broken windows, the police, and neighborhood safety, *Atlant. Mon.* March.
- Wright, R., Tekin, E., Topalli, V., McClellan, C., Dickinson, T., 2014. Less cash, less crime: evidence from the electronic benefit transfer program. NBER Working paper 19996.
- Wright, R.T., Decker, S.H., 1997. *Armed Robbers In Action: Stickups and Street Culture*. Northeastern.
- Yinger, J., 1986. Measuring racial discrimination with fair housing audits: caught in the act. *Am. Econ. Rev.* 76, 881–893.
- Zimring, F., Hawkins, G., Kamin, S., 2001. *Punishment and Democracy: Three Strikes and You’re Out in California*. Oxford University Press, New York, NY.
- Zimring, F.E., 2007. *The Great American Crime Decline*. Oxford University Press, New York, NY.
- Zimring, F.E., 2011. *The City That Became Safe: New York’s Lessons for Urban Crime and Its Control*. Oxford University Press.

This page intentionally left blank

INDEX

Note: Page numbers followed by *f* indicate figures and *t* indicate tables.

A

- Abortion legalization, impact on urban outcomes
 - characteristics changed crime, 1527
 - feminism, 1528
 - legalizing, 1527
 - naïve theory, 1527
 - noncriminal outcomes, 1528
- Access to public transit
 - and age, 1061, 1061*f*
 - American Community Survey (ACS) data, 1061, 1061*f*
 - fixed costs, 1062
 - good/bad access, 1061
 - metro areas, 1061
- ACS. *See* American Community Survey (ACS)
- Adjustable-rate mortgages (ARMs), 775, 785–786
- Agglomeration economies
 - conditional logit models, 1174
 - decentralized tax setting, 1173–1174
 - economic activity, 1207, 1215
 - efficiency-related argument, 1206–1207
 - empirical approach, 1216
 - endogeneity, 1216
 - equilibrium tax competition, 1178
 - estimation strategy, 1217
 - financial incentives, 1174–1175
 - fiscal policy, 1171
 - local-level tax policies, 1176–1178
 - metropolitan economic status, 1096–1097
 - mobility-reducing effect, 1172
 - population sorting, 1173
 - productivity distribution, 1216
 - redistributing activity, 1207–1208
 - regression model, 1175
 - tax competition models, 1172
- Aggregate commuting cost, 996–997
- Agricultural commodities, 1389
- Alonso–Muth–Mills (AMM) model, 1058–1059, 1060, 1060*f*, 1062, 1076–1077
- Amenities
 - and labor supply, 1102–1104
 - location-specific, 1074
 - neighborhood, 906–907
 - nondurable, 1065–1066
 - nontradeable, 1357
 - physical, 1064–1065
 - urban, 1211–1212
- American Community Survey (ACS)
 - access to public transit, 1061, 1061*f*
 - age of housing stock and public transit, work trips, 1061, 1067, 1111*f*
 - pooled census tract data, 1050*f*, 1054–1055
 - state-level and federal programs, US, 1238
- American crime decline
 - of 1991–2001, 1599
 - description, 1586–1587
 - nontraditional (*see* Crime decline, American)
 - traditional (*see* Crime decline, American)
 - twenty-first century, 1599–1602, 1600*t*
- American crime rise, 1965–1975, 1602–1604, 1603*t*, 1604*t*
- American Housing Survey (AHS), 1059–1060, 1078–1080, 1087
- Anti-discrimination laws, Post-Civil Rights era, 1006
- Appalachian Regional Commission, 1202, 1272–1273
- Armington model, 1391
- Asset pricing and housing
 - representative agent model, 787–790
 - risk sharing, housing collateral, 790–792
 - stochastic discount factors, 787
- Auctions
 - bargaining, 845
 - bidding wars, 848–850
 - buyer's participation, 846–847
 - competition and prices, 847
 - condominiums, 845
 - critical time, 846–847
 - dispersion encourages buyers, 847
 - English auctions, 845–846
 - jump bids, 847

Auctions (*Continued*)

- real estate, 847
- transactions, 846

Autonomous local tax share, 1135–1136

B

Bankruptcy Abuse Prevention and Consumer Protection Act, 970–971

Bargaining and directed search, 842–843

Basic tax competition model

- average tax rate, 1147, 1148*f*
- Keen–Kotsogiannis model, 1152
- public consumption, 1148–1149
- tax revenue, 1148–1149

Bidding wars

- auction and sequential search, 849
- bidder chooses, 849
- buyers, 848
- definition, 846
- internet adoption, 849
- listing price, 849
- multifamily residential market, nondistressed, 848
- nondistressed single-family houses, 848
- popularity, 849

BIDs. *See* Business improvement districts (BIDs)

Bilateral commuting flows, 1127, 1128*f*, 1392

Bounded rationality and housing markets

- cognitive limitations, 737–741
- exogenous and heterogeneous beliefs, 733–734
- extrapolative beliefs, 734–737
- normative analysis, 743
- social learning and entrepreneurs, 741–743

Broker-listed *vs.* for sale by owner (FSBO)
properties, 870–871

Broker-owned *vs.* client-owned properties, 869–870

Brueckner–Selod model

- budget constraint, 1001
- defensive expenditures, 1000–1001
- equilibrium solution, 1001–1002
- eviction condition, 1000–1001
- formal residency, 1001–1002
- land consumption, 1000–1001
- Pareto improvement, 1001–1002
- squeezing process, 1001

Bubbles

- housing (*see* Housing bubbles)
- and public policy, 743–746

Buildable land and historical density patterns

aerial photographs, 1311–1312

empirical strategy, 1311

homeowners and developers, 1310

metropolitan areas, 1310–1311

regulation, 1311

Wharton Residential Land Use Regulation Index
(WRLURI), 1312

Building occupancy and land-use certificate
(BOLUC), 999

Business cycle and housing

- Davis and Heathcote model variable properties,
772–773, 772*t*
- house prices and residential investment, 773
- intermediate goods, 770–772
- land acts, 773
- lead-lag relationship, 767
- offset price effects, 773
- real business cycle (RBC) model, 767, 768
- relative prices, 768, 769*f*
- risk shocks, 773
- two-sector economy, 768–770

Business improvement districts (BIDs), 1555

Buyer–seller problem

- calibrated search model, 830
- cyclical properties, housing markets, 831
- houses and households, 829–830
- in-home search, 829
- match quality, 830–831
- Wheaton's model, 829–830

Buyer-side commission rate, 858

C

Calibration theory models, 1179

California enterprise zone program

- Department of Housing and Community
Development, 1231
- economic development, 1236
- evidence, 1232, 1233*t*, 1237
- geographically targeted policies, 1232
- heterogeneity programs, 1232–1236
- job-creating effects, 1236
- targeted employment area (TEA), 1230–1231

Capital employment, 1146–1147

The Castle doctrine effect, 1562

Census tract relative income

- metropolitan statistical areas (MSAs), 1051–1052,
1054–1055, 1074, 1108*f*
- transition rates, 1050–1051, 1051*f*

- Central business district (CBD)
 - commuting costs, 1158–1159, 1160
 - employment, 1160–1161
 - municipalities' employment, 1158–1159
 - suburban governments, 1160
- Central municipalities
 - capital taxes, 1156–1157
 - city's population, 1160
 - decentralization, 1159
 - decision making, 1156
 - income distribution, 1165–1166
 - nondistortionary tax, 1157
 - suburban tax rates, 1156
 - tax policy, 1156
 - uniform tax rates, 1164
- Certainty and severity, 1537–1539
- Churning process
 - description, 1501
 - Grossman–Helpman quality ladder model, 1504–1505
- City dynamics
 - metropolitan economic status (*see* Metropolitan economic status)
 - size distribution of cities (*see* Size distribution of cities)
- City-level taxation
 - agglomeration, 1125
 - central jurisdiction, 1125
 - functional urban area, 1178
 - interdependencies, 1125
- City-size distribution
 - growth, 1499–1500
 - stability, larger cities, 1500–1501
- City's population, 1147
- Climate-inducing migration, 1407–1408
- Clusters policies
 - businesses operating, 1263
 - evidence, 1267–1268
 - firms' innovation productivity, 1263
 - formation and development, 1263–1264
 - innovation process, 1264
 - local productive systems (LPS), 1261–1262
 - regression discontinuity (RD) expenditures, 1263
 - total factor productivity (TFP), 1262
- Cobb–Douglas case, 788
- Commodity flow survey area, 1375
- Community development
 - enterprise zones, 1279
 - estimated effect, 1278
 - local autonomy, 1278
 - New Markets Tax Credit (NMTC), 1275–1276
 - Patti Territoriali program, 1276–1277
 - tax increment financing (TIF), 1275
 - Texas Enterprise Zone Program, 1276
- Community Reinvestment Act of 1977 (CRA), 743, 1007–1008
- Commuting costs
 - decentralization, 1374
 - household expenditure, 1353
 - trade literature, 1352–1353
 - transportation technology, 1363
 - urban literature, 1352–1353
- Conditional logit model, 1168
- Congestion costs, 1391
- Constant elasticity of substitution (CES), 775
- Consumer market access, 1359
- Consumption index, 1356–1357
- Consumption taxes
 - income tax revenue, 1180–1184
 - total tax revenue, 1180–1184
- Corporate income tax revenue, 1180–1184
- Counterfactual, 1169, 1365
- Courant's model, 820
- Crime
 - description, 1521
 - developed countries, 21st century, 1602
 - government policies, 1523
 - historical swings, 1584–1585, 1584f
 - objectives and policy, 1522–1523
 - output/employment, macroeconomists, 1583
 - physical features, 1523
 - policy implications, 1523
 - race and ethnicity, 1534–1536
 - rates of change, crimes per capita in 2000s, 1585, 1586t
 - sodomy and blasphemy, 1521
 - strong claim of criminogenic characteristics, 1525–1528
 - weak claim of criminogenic characteristics, 1528–1534
 - white-collar crimes, 1521–1522
- Crime decline, American
 - abortion, 1590
 - capital punishment, 1590
 - crack, 1590–1591
 - demographics, 1588

- Crime decline, American (*Continued*)
 - electronic banking, 1597–1598
 - in utero* exposure, tobacco and alcohol, 1594–1595
 - interactions, 1595
 - lead gasoline, 1593–1594
 - online shopping, 1598–1599
 - police force size, 1587
 - police productivity, 1591–1592
 - prison population, 1587
 - private protection, 1596–1597
 - prosperity, 1588–1589
 - psychopharmaceuticals, 1595
 - residual, 1589, 1589*t*
- Crime deterrence
 - actual arrest rates, 1541–1543
 - age of majority, 1545
 - California sentence enhancement, 1546
 - collective pardon, 1545
 - hot spots, 1540–1541
 - number of police officers, 1544–1545
 - perceived probability, arrest, 1541
 - prison conditions, 1545
 - probation sanctions, 1543–1544
 - responses to terrorist attacks, 1540
 - and selection, 1555–1556
 - violence reduction initiatives, 1546–1547
- Cross-fertilization, externalities and sorting of industries, 1498–1499
- D**
- Davis and Heathcote model variable, 772–773, 772*t*
- Decentralization of economic activity
 - changing specialization in China, 1501, 1502*t*
 - in China, 1503, 1503*t*
 - manufacturing and centralization of services in India, 1503, 1504*f*
 - modeling industry movement across cities, 1503–1506
 - movement across urban hierarchy, 1506
 - in South Korea, 1501, 1502*t*
- Decentralization theorem
 - autonomous tax, 1136–1137
 - government tax revenue, 1134–1135, 1135*f*
 - instrumentation strategy, 1373–1374
 - interstate highway, 1372–1373
 - jurisdictional fragmentation, 1126
 - local municipal taxes, 1135–1137, 1136*t*, 1137*f*
 - metropolitan statistical areas (MSAs), 1372–1373
 - political and institutional, 1126
 - population and employment, 1373
 - tax authority, 1126
- Developed-country transportation, 1353
- Discretionary policies
 - agglomeration externalities, 1227–1228, 1258
 - asymmetrical effects, 1258–1259
 - capital investment and employment, 1256–1257
 - capital investment subsidy program, 1228
 - characteristics, 1226
 - data aggregation, 1227
 - endogenous, 1227
 - EU Structural Funds, 1255
 - evaluating effects, 1252–1253
 - evidence, 1254, 1259–1261, 1260*t*
 - heterogeneous, 1259
 - potential location, 1253
 - Regional Selective Assistance (RSA) program, 1226–1227, 1254–1255
 - regression discontinuity (RD), 1229
 - spillover effects, 1257–1258
 - time-invariant characteristics, 1228–1229
 - total factor productivity (TFP) growth, 1256
- Domestic trades, 1363
- Dominant municipality, 1140
- Durable housing, filtering and gentrification
 - aging stock and income stratification, 1067, 1068*f*
 - Chicago and Philadelphia, 1067
 - high- and lower-income communities, 1066
 - metropolitan statistical areas (MSAs), 1067
 - New York and Los Angeles, 1067–1068
 - timing and location of development/redevelopment, 1067
- Dynamic models, spatial variation
 - asymmetrical preferences, 1070–1071
 - Card–Mas–Rothstein (CMR) model, 1071–1072
 - Case–Shiller zip code-level repeat sales house price indexes, 1069–1070
 - durable housing, filtering and gentrification, 1066–1068, 1068*f*
 - Guerrieri–Hartley–Hurst (GHH) model, 1069–1070
 - homeownership, 1072–1073
 - isolated neighborhood model, 1070, 1071
 - micro-based survey data, 1072–1073
 - segregation, 1070
 - social dynamics, 1069–1070

spatial proximity model, 1070
tipping models, 1072

E

Emergency Economic Stabilization Act (EESA) of 2008, 972

Empirical evidence

city sizes, 998–999
counterfactual exercises, 1168
decentralization, 1372–1374
equilibrium tax rates, 1150
indirect utility function, 1167–1168
industrial organizations, 1168
intercity infrastructure, 1378–1383
jurisdiction's population, 1151–1152
land invasion, 998–999
local tax rates, 1151, 1151*t*, 1168
migration equilibrium, 999
municipality's population, 1150
politics, 1342
population sorting, 1170
reverse causality, 1166–1167
rural income risk, 998–999
structural estimation, 1170–1171
variance, 999

Employment

and boost investment, 1226
Colorado's enterprise zones, 1223
enterprise zones, 1225, 1230
estimated models, 1218–1219
evidence, 1214, 1248
network models, 1211
Ohio's program, 1246
and productivity, 1216
residential segregation, 1210

Endowments, transportation costs

commuting technology, 1356
consumption index, 1356–1357
equilibrium population density, 1356
income and expenditure, 1357
varieties, 1356–1357

Enterprise zones

California enterprise zone program, 1230–1237
data and econometric methods, 1230
evidence, 1249–1250
US state-level and federal programs, 1237–1246
Zones Franches Urbaines (ZFU's), 1246–1247

Equilibration mechanism

commuting-cost, 995
edge resident, 995
land consumption, 995
living cost, 995
productivity, 995
urban population, 995

Equilibrium population density, 1356–1357

Equilibrium tax rates, 1151

Equity motivations

enterprise zone, 1212
inframarginal workers, 1212–1213
land prices, 1213
urban economics, 1212

European Social Fund (ESF), 1202–1203, 1268–1269

Exclusive agency *vs.* ERTS brokers, 872–873

Expenditure, public, 1145

Expenditure spillovers, 1126

Experimental housing allowance program (EHAP)

entitlement housing voucher program, 903–904
housing policy debates, 903–904
influential demand experiment research, 903
project-based assistance types, 904
tenant-based assistance, 904

Extrapolative beliefs

backward-looking investors, 735
extrapolative bias, 735
financial markets, 735
Gordon growth formula, 734–735
growth rate, 736
optimistic buyers, 735
pricing formula, 735, 736
sophisticated version, 736–737
strict rationality, 735

F

Federal Housing Enterprises Financial Safety and Soundness Act of 1992, 953

Federal Housing Finance Association, 1085–1086

Federal National Mortgage Association, 944

Federal Reserve Board, 743–744

Federal Trade Commission's (FTC), 858

Filtering of housing

American Housing Survey (AHS) panel, 1084–1085
balanced panel of census tracts, 1084
filtering rates, 1085
and gentrification, 1084–1086

- Filtering of housing (*Continued*)
 - rental and owner-occupied units, 1085–1086
 - simple housing demand function, 1085
 - turnovers, home, 1084–1085
 - Financial portfolio
 - agents, 783–784
 - asset allocation puzzle, 783
 - Epstein–Zin preferences, 784
 - equity participation, 784
 - fixed participation cost, 784
 - geography, 784–785
 - hump-shaped life-cycle profile, 784
 - owner-occupiers, 783–784
 - renters and owners, 783–784
 - risky stocks and bonds, 783
 - Fiscal decentralization
 - Government Finance Statistics (GFS), 1132–1134, 1137
 - government tax revenue, 1134
 - International Monetary Fund (IMF), 1132–1134
 - local municipal taxes, 1136, 1136*t*, 1137, 1138*f*
 - local tax share (LTS), 1134
 - municipal tax revenue, 1134
 - Organization for Economic Cooperation and Development (OECD), 1139–1140
 - property taxes, 1137
 - tax competition, 1135–1136
 - Fiscal equalization, 1146
 - Fixed-rate mortgages (FRMs), 773, 785–786
 - Foreclosure process and mortgage default
 - borrower-friendly foreclosure laws, 968–971
 - distribution, 1034–1037
 - externalities, 1037–1039
 - hedonic price regressions, 1037
 - historical data, 965–966
 - Home Owners' Loan Corporation (HOLC), 966–968
 - multifamily foreclosures, 1037
 - nonjudicial foreclosure process, 968–969
 - spatial concentration, 1038
 - Freight over time, 1345, 1346, 1346*f*, 1347*f*
 - Full-commission brokers *vs.* split-commission brokers, 875–876
 - Functional specialization and diversity, 1493
- G**
- General equilibrium theory
 - agglomeration forces, 1364
 - employment, 1389–1390
 - expenditure, 1364
 - fiscal decentralization, 1161
 - income sorting, 1162
 - infrastructure, 1388
 - intercity/intracity effects, 1388
 - local public budgets, 1161–1162
 - market access, 1387
 - property taxation, 1162–1163
 - public good, 1161–1162
 - roadway growth, 1389–1390
 - transportation investments, 1364
 - General-purpose technology (GPT), 1477
 - Gentrification
 - and filtering, 1084–1086
 - local residents, 1086–1087
 - Gibrat's law, 1472–1473. *See also* Population growth
 - dynamics and Gibrat's law
 - Golden Quadrilateral Project, 1383
 - Goods, transportation costs
 - employment, 1349, 1350*t*
 - gross domestic product (GDP), 1343–1344
 - nontraded services, 1343–1344
 - railroads, 1344–1345, 1344*f*
 - revenue, 1344–1345, 1345*f*
 - travel time cost, 1351
 - Government Finance Statistics (GFS)
 - fiscal decentralization, 1180–1184
 - government tax revenue, 1180
 - individual government finances (IndFin) data, 1184
 - local governments, 1180, 1183*t*, 1185*t*
 - local tax share (LTS), 1180
 - Government response, housing crisis
 - Federal Housing Administration (FHA)
 - Modernization Act, 971
 - GE model, 974
 - Housing Assistance Tax Act (HATA), 971–972
 - housing market, 975
 - pre-Housing Affordable Modification Program (HAMP) mortgage renegotiation rates, 974
 - Government-sponsored enterprises (GSEs), 802
 - Grossman–Helpman quality ladder model, 1504–1505
 - Growth *vs.* reorganization
 - endogeneity, 1370–1371
 - hypothetical network, 1371–1372, 1371*f*, 1384
 - infrastructure, 1383–1384

intercity and intracity regressions, 1371
 railroad construction, 1370–1371
 replications, 1372
 transportation infrastructure, 1370–1371
 Guthrie's model, 1082

H

Harris–Todaro model

employment probability, 1412
 land consumption, 1410
 migration equilibrium, 1410
 nonhousing price, 1410
 rural income, 1410
 urban unemployment, 1410
 wages, 1410

“Haurin” model, 822–823

Head Start, 1529

Hierarchy, urban

benchmark model, 1489
 city sizes and hierarchies, 1487–1492
 cross-fertilization, externalities and sorting of industries, 1498–1499
 dynamics, 1499–1506
 functional specialization and diversity, 1493
 nursery cities and product cycle, 1492
 second generation, 1492–1494
 selection and sorting, 1497
 sorting within and across cities, 1495–1496
 spatial allocation, resources, 1506–1512
 third-generation models, 1494–1499

Higher education

evidence, 1267–1268
 and innovative outcomes, 1267
 labor productivity, 1266
 optimal location, 1267–1268
 regional development, 1266
 substantial expansion, 1266
 and university research, 1280

Hodrick–Prescott filter, 768

Home consumption, 768

Home maintenance and housing capital
 depreciation

American Housing Survey (AHS), 1078–1080
 augmented model, 1078–1080
 declining communities, 1081
 declining house prices, 1080–1081
 gross domestic product (GDP), 1078

investment, US residential fixed assets, 1078,
 1079_t
 sample median values, 1080
 US Bureau of Economic Analysis (BEA), 1078
 Home Mortgage Disclosure Act (HMDA) data,
 1007

Homeowners, developers and local politics

empirical evidence, 1308–1310
 endogeneity, 1308
 equilibrium house prices, 1306–1307
 Fischel's argument, 1307
 growth control measures, 1307–1308
 growth machine, 1309
 housing supply regulation, 1307, 1308
 housing wealth, 1306
 restrict development, 1305
 zoning regulation, 1306

Home Ownership and Equity Protection Act
 (HOEPA) in 1924, 950

Homeownership transitions

credit constraints, 1016
 “housing partnerships”, 1018
 intergenerational transfers, 1017
 liquidity constraints, 1016
 shared equity programs, 1018
 zoning regulations, 1016

Home Owners' Loan Corporation (HOLC) Act of
 1933, 943–944

Home redevelopment, decision making

data from seven municipalities, 1082
 floor space, 1083
 Guthrie's model, 1082
 hedonic models, 1083–1084
 parcel of land, 1081
 property ownership, 1082
 sale prices, 1083
 separated home sales, 1081–1082
 structural probit model, 1081–1082
 structure ages and deteriorates, 1082–1083
 time of sale, 1083–1084

“Hostile attribution bias”, 1531

Hot and cold markets, 829

Household expenditure, 1354_t

House prices

with default, endogenous, 799–800
 with default, exogenous, 794–796
 endogenous, no default, 796–799
 exogenous, no default, 793–794

- Housing affordability
 - construction costs, 1005
 - expenditures, 1005–1006
 - income elasticities, 1005–1006
 - investment, 1439–1443
 - land-management policies, 1443–1447
 - squatting, 1004
- Housing Affordable Modification Program (HAMP), 804
- Housing and portfolio choice
 - calibration and estimation, 778–779
 - choices, 775
 - constraints, 775–777
 - home ownership and housing consumption, 774
 - illiquid, 774
 - indivisible, 774
 - labor market, 774
 - pledgeable capital *vs.* households, 774
 - portfolio literature (*see* Portfolio choice and housing)
 - utility, 774–775
 - wages and prices, 777
- Housing bubbles
 - bounded rationality and housing markets, 732–743
 - economic approaches, 703–704
 - empirical regularities, 715–722
 - extrapolative models, 705
 - limitations, 705
 - linear asset pricing model (LAPM), 703, 705–715
 - markets, 703
 - natural expectations, 705
 - public policy, 743–746
 - rational models, 704
 - rational theory, 722–732
 - spatial heterogeneity, 703
 - US, 702–703
- Housing capital, maintenance and depreciation, 1078–1081
- Housing collateral ratio
 - advantage and disadvantages, 792
 - asset-pricing implications, 790–791
 - asset-pricing predictions, 791–792
 - Cobb–Douglas type, 791
 - Lucas-type trees, 790–791
 - persistent and countercyclical movements, 791
 - risk sharing, 791
- Housing consumption
 - choice-based sampling methods, 909–910
 - Current Population Survey (CPS), 908
 - lowest income quintile, 910
 - low-income housing assistance, 906
 - neighborhood amenities, 906–907
 - policy interest, 909
 - project-based housing assistance, 907–908
 - public housing tenants, 907
 - random assignment voucher evaluation, 908
- Housing, cross-sectional facts
 - average household's net worth, 762–763, 764f
 - home ownership rate, 761, 761f, 762, 763f
 - loan to value ratios, 761–762
 - mortgage debt, 761–762, 762f
 - portfolio shares by age, 763–764, 765f
 - saving and lending, 761–762
 - Survey of Consumer Finances (SCF), 762, 763f
- Housing dynamics, empirical regularities
 - bubbles, 720–721, 722
 - excess variance, 715–716, 717f
 - linear asset pricing model (LAPM)/user cost model, 715
 - quantities *vs.* prices, 718–720
 - short run momentum and long run mean reversion, 716–718, 719t, 720f
 - US, 722
- Housing investment and negative equity
 - adverse income shock, 1031
 - capitalization, 1030–1031
 - deficiency judgment, 1033
 - income elasticity, 1032
 - investment decision, 1032–1033
 - local public infrastructure, 1034
 - maintenance and improvement expenditures, 1031–1032
 - precautionary savings, 1031
 - price externalities, foreclosures, 1033–1034
 - recourse mortgages, 1033
 - transitory income shocks, 1031–1032
- Housing markets
 - bargaining (*see* Bargaining)
 - description, 815
 - equilibrium models, 817–819
 - fixed supply model, 737–738, 738t
 - homebuyers, 739
 - institutions and economics agents, 816
 - intermediation factors, 816–817, 818f

- market factors, 816–817, 818*f*
 - microstructure, 816–817, 817*f*
 - multiple listing service (MLS), 816
 - naive spatial extrapolation, 740
 - natural expectations models, 737
 - pricing formula, 737, 741
 - rational expectations agents, 738
 - rational pricing and growth equations, 740
 - search (*see* Search)
 - spatial equilibrium logic, 740
 - transactions, 815–816
 - types, 737
 - variable supply, construction costs, 739, 739*t*
 - Housing policy
 - causal inference, 892
 - difference-in-difference framework, 891
 - economists and policy makers, 800–801
 - federal government involvement, mortgage market, 943–955
 - Federal housing policy, 938
 - finance intermediaries, 802
 - financial and housing crisis, 938–939
 - homeownership rate, 940–943
 - Housing Affordable Modification Program (HAMP), 804
 - housing programs, 889
 - instrumental variables (IV) estimator, 891
 - justifications for promoting homeownership, 939–940
 - loan-level data, 889
 - local average treatment effects (LATEs), 891–892
 - low-income housing assistance, 889
 - low-income rental housing policy, 892–937
 - and monetary, 803–804
 - mortgage interest rate deductibility, 801–802
 - public housing stock, 889
 - quasi-experimental methodology, 889, 890
 - randomized controlled trials (RCTs), 890
 - state and federal interventions, 889
 - tax system, 938
 - Housing search, direct
 - and bargaining, 842–843
 - and booms and busts, 844–845
 - ex ante* price commitments, 839
 - labor/commodity search, 839–840
 - and labor markets, 845
 - market efficiency, 839–840
 - posted price impacts search, 839–840
 - structural approaches, 843–844
 - unemployment and vacancies, 839–840
 - Housing search, price
 - commitment, 840–841
 - partial commitment, 841–842
 - signal, 842
 - strategic instrument, 840
 - Housing supply, data collection
 - beyond surveys, 1303
 - building codes, 1296–1297
 - empirical research, 1303–1304
 - geography, regulation and urban development, 1302–1303
 - heterogeneity, 1294
 - indirect measurement, 1295–1296
 - land use controls, 1297–1302
 - local land use regulations, 1303
 - regulation (*see* Regulation)
 - US Geological Survey (USGS), 1302
 - Housing turnover and negative equity
 - American Housing Survey (AHS) data, 1027–1028
 - annual mortgage payments, 1027–1028
 - equity reductions, 1024
 - financial crisis, 1029–1030
 - financial frictions, 1024, 1027–1028
 - HARP and FHA streamline refinance programs, 1030
 - house price indices, 1026
 - mortgage application, 1025–1026
 - mortgage delinquency, 1028
 - mortgage servicing data, 1025–1026
 - owner-occupied housing turnover, 1024–1025
 - time-fixed effects, 1028
- ## I
- Incapacitation
 - plus deterrence effects, 1574–1575
 - strong claim, 1572–1574
 - weak claim, 1572
 - Incarceration
 - drug enforcement and prisoners, 1569
 - growth of per capita state prison population, 1570, 1570*t*
 - incapacitation (*see* Incapacitation)
 - optimal, 1567–1568
 - parole conditions, 1569, 1571
 - postprison effects, 1575–1578
 - prison admissions per crime, 1570, 1571

- Incarceration (*Continued*)
- prison gangs (*see* Prison gangs)
 - race and mass, 1581–1583
 - rehabilitation and retribution, 1567
 - trade-offs, 1580–1581
 - in US, 1568
- Incentive misalignment, real estate brokerage
- asymmetrical information, 865–866
 - broker-listed *vs.* FSBO properties, 870–871
 - broker-owned *vs.* client-owned properties, 869–870
 - broker representation, 873–875
 - competitive equilibrium, 867–868
 - costly search for both buyers and new sellers, 867–868
 - exclusive agency *vs.* ERTS brokers, 872–873
 - full-commission brokers *vs.* split-commission brokers, 875–876
 - hidden action/moral hazard, 865–866
 - hidden type/adverse selection, 865–866
 - misalignment, 866–869
 - mitigating, 877
 - multiple brokers, 867–868
 - multiple listing service (MLS), 868
 - one-seller-one-broker setting, 867
 - one-sided seller search, 868–869
 - percentage commission system, 868
 - principal-agent setting, 865–866
 - in rental markets, 876
 - sellers and agents, 866
 - traditional brokers *vs.* discounted brokers, 871–872
 - Yavas's model, 866–867
- Incentives
- income, earnings and employment, 1547–1550
 - race and, 1550–1552
- Income distribution
- municipality, 1155
 - sorting equilibrium, 1155
- Income growth dynamics, 1467–1468, 1470–1472
- Income segregation in US cities, 1058
- Indirect measurement, data collection, 1295–1296
- Indirect utility function, 1168
- Industrial clusters
- enterprise zone program, 1261
 - evidence, 1265–1266
 - and wider effects, 1264
- Industrialization and urbanization
- agricultural-based, 1460
 - of economic activity, 1461
 - Stone–Geary utility function, 1460
 - total factor productivity (TFP), 1460
- Industry-level agglomeration, 1175
- Industry localization, 1209–1210
- Inelastic housing supply
- description, 1104
 - durable housing in declining cities, 1105–1106
 - regulation in growing cities, 1104–1105
- Infrastructure investment
- Appalachian Regional Commission (ARC), 1272–1273
 - cost–benefit analysis, 1272
 - estimate effects, 1271
 - European Union, 1268–1269
 - federal subsidy program, 1272
 - generosity, 1274
 - manufacturing employment density, 1271
 - Structural Funds program, 1273
 - Tennessee Valley Authority (TVA), 1269, 1275
 - treatment intensity, 1274
- Infrastructure, transportation costs
- economic outcomes, 1368
 - employment growth, 1374–1375
 - ordinary least squares (OLS), 1367–1368
 - political economy, 1387
 - quasi-random variation, 1368
 - variability activities and modes, 1386–1387
- Instrumental variables, 1389–1390
- Intercity infrastructure
- high-income, 1378–1380
 - low-income, 1380–1383
- Interjurisdictional variation, 1611–1612, 1612*t*
- Intermetropolitan variation, 1604–1606, 1605*t*, 1606*t*
- Internal organization, transportation
- endogeneity, 1376–1377
 - population density, 1377
 - postcode unit, 1377
 - public policy interest, 1378
 - regressions, 1376
 - subways, 1376
- International Monetary Fund (IMF), 1132–1134
- International trade, 1346, 1349*t*
- Internet and housing
- approaches, 835
 - buyer and seller liquidity, 836

Carrillo's analysis, 836
 description, 835
 empirical analysis, 835–836
 lower cost search, 835
 model's predictions, 835–836
 ordinary least squares (OLS) empirical analysis, 836
 properties, 836
 reduced-form approaches, 837
 visual information, 836
 Interstate Construction Program, 1340–1341
 Intrametropolitan variation, 1606–1611
 Intraurban tax competition, 1179
 Investments
 foreign and domestic investment, 1004–1005
 gross domestic product (GDP), 1004
 income elasticities, 1004
 informal housing, 1004
 infrastructure and services, 1447–1448
 Organization for Economic Cooperation and Development (OECD), 1004–1005
 supply responseiveness, 996*f*, 1004–1005
 Isolated neighborhood model, 1070, 1071

J

Jimenez model
 budget constraint, 1000
 community, 1000
 comparative-static analysis, 1000
 consumer, 1000
 demand and supply, 1000
 eviction probability, 1000
 formal residence, 1000
 housing expenditure, 1000
 optimization, 1000
 and Turnbull models, 1030–1034
 Jurisdictional fragmentation, urban
 functional areas, 1127, 1128*f*
 government tax revenue, 1139–1140, 1139*f*
 inhabitants, 1132
 local governments, 1139, 1142, 1144*f*
 Organization for Economic Cooperation and Development (OECD)/European Commission (EC), 1127, 1128
 tax competition, 1139–1140
 uniform distribution and rank-size, 1142

K

Knowledge spillovers, 1208–1209

L

Labor earnings and employment
 Aid to Families with Dependent Children (AFDC), 916–917
 labor force participation, 918
 low-income housing program, 917–918
 Low-Income Housing Tax Credit (LIHTC), 922
 mental and physical health, 917
 neighborhood occupied, 919
 nominal benefit reduction rates, 920
 productivity of children, 916
 project-based rental assistance, 921
 PSID-Assisted Housing Database, 922
 standard economic theory, 917
 Welfare to Work Voucher program, 919
 work disincentive effects, 921
 Labor market equilibrium
 commuting technology, 1360
 and direct housing search, 845
 land prices, 1360
 population mobility, 1359
 real income, 1359
 Land-management policies
 colonial planning, 1007
 consumption standards, 1006
 density regulations, 1444–1446
 economic development, 1007
 growth controls, 1006
 slum formation, 1007
 urban growth boundaries (UGBs), 1446–1447
 Land use controls
 Boston metropolitan area, 1301
 data and market coverage, 1301
 development, 1297–1298
 Glaeser-Schuetz-Ward (GSW), 1298, 1301
 Gyourko-Saiz-Summers (GSS), 1298
 heterogeneity, 1298
 historical data, 1302
 house price growth, 1297
 local zoning codes, 1302
 regulations, 1298
 and rental residential houses, 712
 search, endogenous housing supply and acquisition, 733–734

- Land use controls (*Continued*)
 - Wharton Residential Land Use Regulation Index (WRLURI), 1299–1301, 1300*t*
- LBD. *See* Longitudinal business database (LBD)
- Lender Processing Services Applied Analytics database, 970
- Life cycle of industries and spatial distribution, 1477–1481
- Linear asset pricing model
 - elastic price booms, 713, 714
 - endogenous supply, 713
 - eschew rental data, 712
 - housing market, 705, 706, 711
 - idiosyncratic market, 709
 - intertemporal user cost formula, 711
 - momentum, mean reversion and stochastic growth rates, 707–709
 - rationality, 714
 - reflect transaction costs, 710
 - short-selling housing, 712–713
 - stocks and bonds, 712
 - user cost model, 706–707, 714
 - US stock market, 709–710
 - widespread nature, 712
- Liquidity, housing markets
 - adjustment, 833
 - analysis, 833–834
 - buyer and seller reservation prices, 833
 - commercial real estate, 833
 - frictionless rental market, 832
 - Heckman's sample selection methods, 833
 - hedonic coefficients, 833
 - owner-occupied houses, 832
 - and price, 832–833
- Local administrative units, 1127
- Local average treatment effects (LATEs), 891–892
- Local government decisions, 1153
- Local-level tax policies
 - empirical literature, 1176
 - industry-level, 1177
 - public goods, 1177
 - reverse-causation, 1176
 - subsidies, 1177
 - Swiss metropolitan areas, 1177
- Local municipal revenue, 1180, 1187*t*
- Local planning authorities (LPAs), 1321
- Local productive systems (LPS), 1261–1262, 1268
- Local public services
 - large-scale voucher program, Chicago, 1064
 - limitation, 1062
 - peer effects, 1063–1064
 - process types, equilibrium, 1063
 - reform, 1064
- Local tax revenue, 1134–1135, 1135*f*, 1137, 1137*f*
- Logistics networks, 1341
- Longitudinal business database (LBD), 1240–1241
- Low-income rental housing policy
 - aggregate consumption, 935
 - benefits distribution, 915–916
 - children's educational outcomes, 923–925
 - consumption distortions, 936
 - cost-effectiveness, 904–906
 - design, housing policy reforms, 937
 - discretionary expenditures, 936
 - goods consumption, 913–914
 - government program, 895
 - HOME Investment Partnerships Program, 936
 - household composition, 926
 - housing consumption, 906–910
 - justification, 893–894
 - labor earnings and employment, 916–922
 - low-income families, 894
 - market prices, 932–933
 - mean tenant benefit and subsidy, 914–915
 - minimum housing standards, 894
 - neighborhood amenities, 910–912
 - neighbors, assisted households, 926–932
 - private provision, 933–934
 - racial and economic integration, 912–913
 - recipient health, 925–926
 - types, 892–893
 - welfare programs participation, 923
- Low-income, transportation
 - agricultural land, 1380–1381
 - gross domestic product (GDP), 1381–1382
 - highway network, 1383
 - hypothetical network, 1381–1382
 - intercity regression, 1380
 - populations, 1381
 - railroad, 1380
- LPAs. *See* Local planning authorities (LPAs)
- LPS. *See* Local productive systems (LPS)
- Lucas-Prescott-style job mobility, 781

M

- Macroeconomics and housing
 - advantages, 754
 - and asset pricing, 787–792

- boom and bust (*see* Real estate)
- and business cycle, 767–773
- and financial crisis, 754
- and macroeconomy, 754
- policy (*see* Housing policy)
- stylized facts, 756–767
- Major social experiments
 - Experimental Housing Allowance Program's (EHAP), 903–904
 - Moving to Opportunity (MTO), fair housing demonstration program, 899–902
 - Welfare to Work Voucher evaluation, 897–899
- Manufacturing *vs.* services, 1476–1477, 1476*t*, 1477*f*
- Market wages, 1358–1359
- Mean reversion, neighborhood economic status
 - census tracts, 1057
 - Occupational Score Index (OCCSCORE)
 - measure, 1056–1057
 - panel unit root tests, 1057–1058
 - period length, 1057
 - Philadelphia County regressions, 1057
 - serial correlation in growth, 1056–1057, 1056*t*
- Mental illness, 1526
- Metropolitan economic status
 - agglomeration economies, 1096–1097
 - amenities and labor supply, 1102–1104
 - culture, 1097
 - inelastic housing supply, 1104–1106
 - natural advantages, 1094–1096
 - productivity and labor demand, 1098–1102
- Metropolitan statistical areas (MSAs)
 - census tract median income relative, 1049–1050, 1050*f*
 - description, 781, 1369
 - fixed effects, 1074–1076, 1075*t*, 1077
 - level, 821
 - and neighborhood, 1049
- Migration equilibrium, 1409
- Monetary and housing policy, 803–804
- Mortgage choice
 - fixed-rate mortgages (FRMs)/adjustable-rate mortgages (ARMs), 785–786
 - and housing, 786–787
 - portfolio choice and life-cycle literature, 786–787
- Mortgage credit
 - access, 1006–1008
 - high-cost lending distribution, 1008–1015
 - over life cycle, 1015–1021
- Mortgage finance, demand shifts
 - abandonment, housing market demand and supply, 1001–1002, 1002*f*
 - down payment constraint, 1004
 - housing's durability, 1000
 - intrametropolitan effects, 1004–1005
 - market-clearing quantities, 1000
 - price and quantity dynamics, 1004–1005
 - residential investment activity cycle, 1002–1003
 - securitization, 1005
 - spatial organization, activities, 1004
 - tax reformers, 1003
- Mortgage finance, United States
 - assumability and portability, 994
 - collateralized loans, 991–993
 - demand shifts, price and quantity, 1000–1005
 - demand volatility, 988
 - design features, 991–995
 - evolution, mortgage contract, 989
 - home equity, 995–997
 - homeownership rate, 989, 990*f*
 - housing investments, 994–995
 - Lien priority, 993
 - neighborhood housing prices, 989
 - nonprime credit, 989
 - nonrecourse/recourse, 994
 - owner-occupied housing units, 989, 991*f*, 997–1005
 - retirement and precautionary saving, 1039–1040
- Mortgage lending
 - empirical strategy, 1013
 - endogenous switching framework, 1011–1012
 - Federal Housing Administration (FHA) fixed-rate home purchase loans, 1013–1014
 - financial literacy, 1008–1009
 - home equity, 1020
 - Home Equity Conversion Mortgage (HECM), 1020–1021
 - homeownership and mortgage, 1015–1016, 1015*f*
 - housing appreciation, 1009–1010, 1019
 - housing debt, 1019
 - instruments and institutions, 1018
 - maintenance and home investment, 1021
 - matched-pair field experiment, 1014
 - pricing differentials, 1013
 - pricing disparity, 1012
 - prime and subprime markets, 1012
 - residential real estate market, 1014–1015
 - reverse mortgage market, 1021
 - standard economic theory, 1015

- Mortgage lending (*Continued*)
 subprime market, 1008–1009, 1008*f*
 subprime mortgages, 1009
 transitions, homeownership, 1016–1018
- Mortgage market, federal government involvement
 affordable housing goals, 959–963
 Federal Housing Administration (FHA), 945
 financial crisis, 943
 government intervention, 944–945
 government role, mortgage default and
 foreclosure process, 965–971
 government-sponsored enterprises
 (GSEs)–securitized loans, 963–965
 home equity, 950
 Home Ownership and Equity Protection Act
 (HOEPA), 950
 low-income and minority homeownership
 policy, 951–954
 postwar surge in homeownership, 955–958
 rise, homeownership, 958–959
 securitization, 954–955
 subprime purchase loans, 950
 US homeownership rate, 1890–2012, 940, 941*f*
- Moving to Opportunity (MTO)
 low-income housing policy, 899
 low poverty voucher group, 901
 types, housing assistance, 900, 901*t*
- Multijurisdiction models
 calibration and estimation, 1162
 housing prices, 1161–1162
 income taxation, 1163–1164
 local property taxation, 1161
- Multinomial logit model, 1403
- Multinomial response framework, 1167
- Multiregion model, 1341–1342, 1355–1356
- Municipalities
 equilibrium tax rates, 1149
 fragmentation, 1130–1132
 jurisdictional fragmentation, 1132, 1133*f*
 logarithm, population, 1129
 Organization for Economic Cooperation and
 Development (OECD), 1130–1132, 1131*f*
 population shares, 1140, 1141–1142, 1141*t*, 1147
- N**
- National Association of Realtors' (NAR) buyers and
 sellers surveys, 821
- The National Crime Victimization Survey (NCVS),
 1585
- National highway system, 1372
- National Housing Act (NHA) of 1934, 944
- National Longitudinal Survey of Youth (NLSY79),
 1026
- National Trunk Highway System, 1340–1341
- NCVS. *See* The National Crime Victimization
 Survey (NCVS)
- Negative equity
 fraction of mortgages, 1022, 1022*f*
 and housing investment, 1030–1034
 housing investment expenditures, 1022–1023
 and housing turnover, 1024–1030
 price discovery, 1022–1023
 and strategic default, 1023
- Neighborhood economic status
 abandoned and rundown buildings, 1052–1053
 access to public transit, 1074, 1076
 Alonso–Muth–Mills (AMM) model, 1074,
 1076–1077
 census tract relative income (*see* Census tract
 relative income)
 changes, 1055–1056
 Detroit's population, 1053–1054
 dwelling age variables, 1076
 filtering, 1084–1086
 gentrification, 1084–1087
 Harlem, 1052
 home maintenance and depreciation, 1078–1081,
 1079*t*
 income displays sharp and nonrandom spatial
 patterns, 1049–1050
 income segregation, US cities, 1058
 local public services, 1074
 location-specific amenities, 1074
 mean reversion, 1056–1058
 MSA (*see* Metropolitan statistical areas (MSAs))
 pace and timing of change, 1077–1086
 public use microdata area (PUMA), 1051–1052
 rich suburbs and poor cities, 1054–1055
 spatial variation, 1058–1073
 timing and location of development and
 redevelopment of housing stock, 1074
 US context *vs.* Europe, 1088
 US metropolitan areas, 1048–1049
- Network effects
 advantages, 1250–1251
 evidence, 1220
 excess clustering, 1220
 Hispanic job density, 1220–1221

- JobsPlus program, 1251–1252
- labor markets, 1211, 1219–1220
- place-based policies, 1221
- spatial mismatch, 1211
- urban amenities, 1211–1212
- New economic geography models, 1179
- New Markets Tax Credit (NMTC), 1275–1276, 1279
- Nondurable amenities, 1065–1066
- Nontradeable amenities, 1357
- Nursery cities and product cycle, 1492

O

- One-sided buyer search model
 - building blocks and motivation, 820
 - Courant's model, 820
 - discrimination, 820–821
 - empirical research, 821
 - ex ante* distribution, 820
 - welfare economics, 820–821
- One-sided search
 - asking price, 824
 - buyer, 820–821
 - characterizing, 820
 - duration, 822–823
 - market factors, 819
 - seller equity, 823
 - seller motivation, 823
 - Simon's analysis, 820
- Optimal city, urbanization
 - agglomeration economies, 997–998
 - commuting cost, 997
 - density externality, 997
 - efficiency, 997
 - equilibrium, 991*f*, 998
 - interjurisdictional transfers, 998
 - land consumption, 996–997
 - marginal product, 998
 - productivity, 998
 - public economics, 998
- Organization for Economic Cooperation and Development (OECD)
 - convergence, 1124
 - jurisdictional fragmentation, 1179, 1180, 1182*t*
 - local governments, 1128–1129, 1181*t*
 - municipalities, 1129, 1129*f*, 1178
 - population clusters, 1128
 - tax revenue, 1124

- Owner-occupied housing
 - down payment requirement, 997
 - federally taxable income, 998–999
 - housing demand curve shifters, 999
 - implicit rental income, 998
 - liquidity constrained households, 999–1000
 - personal bankruptcy proceedings, 998
 - public policy, 997–998
 - tax code, 999

P

- Panel study of income dynamics (PSID), 778–779, 1017, 1086–1087
- Panel unit root tests, 1057–1058
- Pareto distribution, 1155
- Peer effects and strategic complementarity, 1559–1561
- Physical amenities, 1064–1065
- Place-based policies
 - agglomeration economies, 1206–1208, 1215–1217
 - disadvantages, 1199
 - discretionary effects, 1226–1229
 - displacement effects, 1225–1226
 - economic activity, 1279
 - empirical analysis, 1213–1215
 - enterprise zone program, 1199, 1201, 1202, 1280
 - equity motivations, 1212–1213
 - EU legislation, 1203
 - European Social Fund (ESF), 1202–1203
 - geographic targeting, 1222–1224
 - government efforts, 1198
 - industry localization, 1209–1210
 - infrastructure expenditure, 1280
 - knowledge spillovers and economy, 1208–1209
 - labor economics literature, 1199–1200
 - labor-market activity, 1281–1282
 - labor mobility, 1206
 - measuring local areas and economic outcomes, 1222
 - multiple interventions, 1225
 - network effects, 1211–1212, 1219–1221
 - policy interventions, impacts, 1230–1279
 - political/fiscal decentralization, 1200–1201
 - relative *vs.* absolute effects, 1229
 - research evidence, 1200
 - spatial mismatch, 1210, 1217–1219
 - strategies, 1198–1199

- Place-based policies (*Continued*)
 - theoretical motivations and behavioral hypotheses, 1215–1221
 - types, 1203, 1204*t*
 - Policy interventions, impacts
 - clusters, 1261–1264
 - community development, 1275–1279
 - discretionary grant-based policies, 1252–1261
 - enterprise zones, 1230–1250
 - infrastructure investment, 1268–1275
 - network effects, 1250–1252
 - universities, 1264–1267
 - Population density, 1127, 1465–1467, 1466*f*, 1467*f*
 - Population growth dynamics and Gibrat's law
 - agricultural/nonagricultural goods, 1468
 - convergence, 1468
 - density, structural transformation, 1468
 - locations distribution, 1464
 - from 1800 to 2000 for US, 1464, 1466*f*
 - from 1880 to 2000 for US, 1464, 1464*f*
 - Population sorting, urban
 - central business district (CBD), 1165
 - commuting cost, 1165
 - general equilibrium models, 1161–1164
 - income elasticity, 1165–1166
 - local governments, 1142, 1190*t*
 - welfare, 1166
 - Portfolio choice and housing
 - financial portfolio, 783–785
 - home ownership, 780–781
 - life-cycle profiles, 780
 - mortgage choice, 785–787
 - price risk and demands, 783
 - smooth consumption, 781–783
 - Price competition, residential real estate brokerage industry, 858–859
 - Prison gangs
 - gang membership, 1578–1579
 - preprison effects and prison culture, 1579–1580
 - state-sponsored deterrence, 1578
 - Probation, 1543–1544
 - Production patterns, urban hierarchy
 - relative industry composition position, 1484–1485, 1484*f*, 1485*f*
 - specialization, 1486–1487, 1486*t*
 - standardized services, 1483
 - Productivity and labor demand
 - central place theory, 1098
 - Duranton model, 1099
 - industries, 1099, 1100
 - innovation and growth, 1099
 - long-run population growth, 1101
 - mobile industries, 1100–1101
 - MSA-level *per capita*, 1101–1102
 - Nursery City paper, 1101
 - probability of innovation, 1100–1101
 - research firms, 1099
 - steady state, 1101
 - Property taxation
 - local government revenue, 1163–1164
 - Tiebout-type equilibrium, 1162–1163
 - Psychopharmaceuticals, 1595
 - Public consumption goods, 1152–1153, 1158
 - Public intermediate investments, 1149
 - Public policy and bubbles
 - asset values, 745
 - Community Reinvestment Act, 743
 - federal government, 745
 - Federal Reserve Board, 743–744
 - housing supply, 746
 - land use restrictions, 746
 - macroeconomic policy making, 744
 - pro-home borrowing policies, 745
 - real estate prices, 744
 - Public services
 - development policies, 1406
 - municipality level, 1405–1406
 - productivity benefits, 1405–1406
 - wage differences, 1405–1406
 - water and electricity, 1406
 - Public transit tends, 1060–1062, 1061*f*
 - Public use microdata area (PUMA), 1051–1052
- ## Q
- Queuing model, 859
- ## R
- Race and ethnicity, 1534–1536
 - Race and mass incarceration, 1581–1583
 - Racial residential segregation, 1557–1559
 - Railroads
 - Fogel's classic study, 1342
 - shipments, 1349–1350
 - Randomized controlled trials (RCTs), 1529
 - Random matching models, housing search bargaining, 825

- behavioral issues, 834
- buyer-seller problem, 829–831
- cycles in housing markets, 825–826
- demand shock, 827–828
- elements, 826–827
- entry, 829
- excess volatility, 825–826
- idiosyncratic match utility, 827
- intensity choice, 828
- labor economics, 825
- liquidity (*see* Liquidity, housing markets)
- National Association of Realtors' buyers and sellers surveys, 828
- opportunistic matching, 831
- price-volume correlation, 827
- seasonality and market thickness, 831–832
- Rational bubbles
 - CEO, 730
 - deterministic, 729
 - endogenous supply, 730
 - extreme case, 729
 - extreme volatility, 732
 - housing prices, 730
 - idiosyncratic risk, 731
 - mortgage insurers, 731
 - mortgage process, 728–729
 - option pricing model, 731–732
 - standard transversality condition, 729–730
 - stochastic, 729
- Rational theory
 - agency and bubbles, 728–732
 - credit markets, 727, 732
 - dynamic model, 727
 - elastic supply, 726
 - heterogeneity, 726
 - housing price changes, 724–725
 - housing prices, 728
 - liquidity asset pricing model (LAPM), 725, 726
 - mortgage approval, 727
 - search, learning and momentum, 722–724
- RCTs. *See* Randomized controlled trials (RCTs)
- RD. *See* Regression discontinuity (RD)
- Real business cycle (RBC) models, 767, 768
- Real estate
 - agent framework, 792
 - and asset prices, 711
 - booms and busts, 702–703, 744
 - bubbles, 702, 703
 - and business cycle, 767–773
 - empirical regularities, 715
 - endogenous house prices, 796–800
 - exogenous house prices, 793–796
 - and housing economics, 703
 - public policy, 745
 - rational models, 730
 - selling stock, 742
- Real estate brokerage
 - collusive behavior, 862
 - collusive pricing, 860
 - competitive pressures, 862
 - Consumer Expenditure Survey (CEX), 862
 - fixed-percentage commission rate, 859, 860
 - flat-fee agents, 862
 - market forces and institutional factors, 861–862
 - multiple listing service (MLS), 860–861
 - North American residential markets, 861
 - queuing model, 859
 - uniform fee across clients, 859
 - virtual office websites (VOW), 861
- Real estate brokers
 - activities, 850–851
 - commission fees, 854
 - competition, residential (*see* Residential real estate brokerage industry)
 - description, 850
 - excessive intermediation, 854
 - homebuyer/seller, 853–854
 - imperfect information, 852
 - incentive issues, 865–877
 - information advantage, 854
 - intrinsic characteristics of housing, 852–853
 - market structure characterization, 854
 - multiple listing service (MLS), 853
 - online resources, 853
 - types of uncertainty, 852
- Regional selective assistance (RSA), 1226–1227, 1253–1255, 1257–1258
- Regression discontinuity (RD)
 - evaluation strategy, 1273
 - exogenous variation, 957, 961
 - and Low-Income Housing Tax Credit (LIHTC), 928
 - treatment, 891

Regulation effects

- beyond housing markets, 1325–1327
- buildable land and historical density patterns, 1310–1312
- cross-sectional evidence, 1292–1293
- data collection, 1294–1304
- determinants, 1304–1316
- dynamic structural model, 1321–1322
- empirical work, 1320, 1322, 1330, 1331
- estimated effects, 1317
- government intervention, 1293–1294
- heterogeneity, 1290
- homeowners, developers and local politics, 1305–1310
- house price cycle and metropolitan areas, 1322
- house prices, 1318
- housing supply, 1293, 1316, 1319, 1321, 1330–1331
- land values and zoning, 1304–1305
- local land use, 1292
- local planning authorities (LPAs), 1321
- numerous papers, 1317
- panel data, 1319–1320
- predicted effects, 1316–1317
- price and quantity, housing, 1316–1322
- real construction costs and house prices, 1290–1291, 1291*f*
- residential building permits, 1318
- residential development, 1291–1292, 1320
- residential permits, 1319
- restrict land use, 1290
- sorting and strategic interactions, 1312–1316
- urban form and homeownership, 1322–1325
- volatile and construction, 1321
- welfare implications, 1327–1330
- zoning laws, 1292

Remoteness cost, 1481–1482

Representative agent model

- adjustment costs, 789–790
- Cobb–Douglas case, 788, 789
- Epstein–Zin preferences, 789
- Lagrange multiplier, 788
- logarithm of SDF, 788–789
- nonconvex adjustment costs, 789–790
- nonhousing goods and services, 787–788
- rental price of one unit, 788–789
- standard Lucas–Breedon kernel, 789
- stochastic discount factors (SDF), 789–790

yields housing services, 787–788

Residential real estate brokerage industry

- entry, 855–857
- National Association of Realtors' (NAR), 855
- price, 858–859
- product differentiation, 857
- Rich suburbs and poor cities, 1054–1055
- Risk sharing, housing collateral, 790–792
- Road congestion, fundamental law, 1375–1376
- Road networks, 1375–1376
- Route instrumental variable, 1368–1369
- RSA. *See* Regional selective assistance (RSA)
- Ruralization *vs.* suburbanization, 1481
- Rural–urban migration
 - city growth, 1409
 - covariation, 1405
 - description, 1459–1460
 - economic growth, 1402
 - empirical evidence, 1404
 - Harris–Todaro model, 1409
 - high birth rates, 1402
 - income risk, 1408–1409
 - industrialization and urbanization, 1460–1461
 - internal migration, 1402
 - middle-income, 1403
 - public services, 1405–1407
 - and transition to modern growth, 1461–1462
 - unemployment rate, 1404
 - without industrialization, 1462–1463

S

Salant's model, 822

Schelling's models, 1070

Search

- directed, 839–845
- focused, 837–838
- one-sided, 819–825
- presearch, 835
- random matching models, 825–835
- segmented, 838–839

Sectoral employment growth, US counties,

1476–1477, 1477*f*, 1479, 1480*f*

Shah model

- eviction costs, 1002–1003
- formal residents, 1003
- government-owned land, 1002
- hybrid model, 1003
- land consumption, 1002

- Shipping cost, 1349–1350
- SIC. *See* Standard Industrial Classification (SIC) subsectors
- Simon's analysis, 820
- Size distribution of cities
 - fast changes, 1090–1091, 1091*t*
 - rank-size rule, 1089, 1090*t*
 - slow movement, 1090
 - very fast movement, 1091–1094, 1093*f*
- Social inefficiency
 - average price of land in city increases, 863–864
 - brokerage cost, 864
 - business-stealing effect, 865
 - Census of Population and Housing, 863–865
 - foregone income, 864
 - low-cost programs, 865
 - rebate bans, 864–865
 - theoretical work, 863
- Social interactions, 1164
- Sorting and strategic interactions
 - characteristics, 1314
 - empirical evidence, 1315
 - fiscal capacity, 1313–1314
 - geographic constraints, 1315–1316
 - land use policies, 1314–1315
 - potential residents, 1315
 - residential development, 1313
 - zoning restrictions, 1312–1313, 1314
- Spatial allocation
 - and agglomeration, 1468, 1473–1475
 - government policies and institutions, 1506–1507
 - locations, 1463
 - manufacturing and services, 1458
 - policy makers, 1458
 - transport investments and technological change, 1508
 - urban and political city bias, 1508–1512, 1510*f*
- Spatial mismatch, 1210, 1217–1219
- Spatial proximity model, 1070
- Spatial variation in neighborhood economic status
 - dynamic models, 1066–1073
 - static models, 1058–1066
- Speculation
 - boom and bust, 713
 - optimistic land, 714
 - real estate, 713
- Squatting economics
 - decision-making, 999–1000
 - defensive expenditures, 1003
 - Ecuadorian survey data, 999–1000
 - eviction risk, 1003
 - hedonic price function, 1003
 - informal dwellings, 1004
 - Jimenez model, 1024–1030
 - land invasions, 1003
 - land-titling program, 1003
 - occupancy, 999–1000
 - renter-protection laws, 999–1000
 - return rate, 1003
 - rural-urban migration, 1003
 - slum dwellers, 1004
 - tenure insecurity, 1003
 - transferability, 1003
- Standard Industrial Classification (SIC) subsectors, 1479
- State-level and federal programs, US
 - American Community Survey (ACS), 1238
 - econometric approach, 1239
 - empirical challenges, 1237
 - empowerment zones, 1240–1241, 1243, 1244
 - enterprise zone policies, 1240, 1246
 - equilibrium model, 1242
 - estimated effects, 1240
 - estimate jobs creation, 1242
 - freedman report analyses, 1238
 - growing literature, 1245
 - longitudinal business database (LBD), 1241
 - Ohio's program, 1246
 - overlapping programs, 1239
 - resident employment, 1237–1238
 - spillover effects, 1243–1244
 - Texas program, 1237
 - types, 1239–1240
 - zone designation, 1243
 - zone residents, 1241
- Static models, spatial variation
 - age of housing stock, 1061, 1067, 1111*f*
 - Alonso-Muth-Mills (AMM), 1059, 1060, 1060*f*
 - commuting costs, 1058–1059
 - income stratification, 1059, 1060*f*
 - local public services, 1062–1064
 - metropolitan area, 1058–1059
 - nondurable amenities, 1065–1066
 - ordinary least squares (OLS) estimates, 1059–1060
 - physical amenities, 1064–1065

Static models, spatial variation (*Continued*)
 public transit and access to cars, 1060–1062
 Stochastic discount factors (SDFs)
 asset pricing and housing, 787
 Stone–Geary utility function, 1460
 Strong claim of criminogenic characteristics
 abortion (*see* Abortion)
 age and gender, 1525
 childhood lead exposure, 1525–1526
 description, 1524
 family structure, 1526–1527
 schizophrenia, attention deficit hyperactivity
 disorder (ADHD) and major depressive
 disorder, 1526
 Stylized facts
 cross-sectional facts, 761–764
 Federal Reserve Board's Flow of Funds Accounts
 tables, 756–757
 housing wealth and structures to gross domestic
 product (GDP), ratios, 756–757, 756*f*
 price-rent ratios, 759–760, 759*f*, 760*f*
 residential investment to gross domestic product
 (GDP), ratio, 756–757, 757*f*
 total spending on housing and utilities, ratios,
 758–759, 758*f*
 US data, 756–757
 volatilities and correlations, 764–767, 766*t*
 Suburban tax policy, 1157
 Survey of Consumer Finances (SCF), 762, 763*f*,
 1026–1027
 Survey of Income and Program Participation (SIPP),
 1016–1017

T

Targeted employment area (TEA), 1201,
 1230–1231
 Tax competition
 autonomy, 1124
 capital employment, 1146–1147
 fiscal decentralization, 1126
 intraurban jurisdictions, 1179
 local municipal governments, 1180, 1189*t*
 local property taxes, 1178
 mobility, capital, 1149
 municipalities, 1125
 Organization for Economic Cooperation and
 Development (OECD), 1127
 production technology, 1147
 rates and population, 1146
 stylized facts, 1127
 urban fiscal competition, 1145
 Tax decentralization, 1140
 Tax-exporting effect, 1158
 Tax increment financing (TIF), 1275
 TEA. *See* Targeted employment area (TEA)
 Tennessee Valley Authority (TVA)
 aggregate effects, 1270
 characteristics, 1269
 development and modernization policy, 1269
 employment growth, 1271–1272
 public infrastructure, 1270
 Tenure insecurity
 building occupancy and land-use certificate
 (BOLUC), 999
 commodity, 999
 continuum, 999
 customary tenure, 999
 idiosyncratic ability, 999
 land allocation, 999
 legal papers, 999
 loan collateral, 999
 slum residents, 999
 squatting, 1023
 urban cost, 999
 Texas Enterprise Zone Program, 1276
 TFP. *See* Total factor productivity (TFP)
 Tiebout equilibrium, 1162–1163
 TIF. *See* Tax increment financing (TIF)
 Total factor productivity (TFP), 1460
 Traditional brokers *vs.* discounted brokers, 871–872
 Transportation costs
 empirical literature, 1366
 employment/land rent, 1366–1367
 endogenous organization, 1351
 geographic distribution, 1366
 gross domestic product (GDP), 1340–1341
 inconsequential units, 1369–1370
 internal combustion engine, 1341
 intracity and intercity regression, 1366
 investments, 1367
 labor shares, 1349
 production technology, 1357–1358
 trade flows, 1360–1361
 Transportation, high-income
 interstate highway system, 1378, 1379
 manufactured goods, 1379

- outcome measures, 1378
 - roads, 1380
 - wages, 1379
 - Transportation, stylization facts
 - commuting costs, 1343
 - external costs, 1355
 - goods, 1343–1352
 - gross domestic product (GDP), 1343–1344
 - household travel, 1352–1354
 - Turnbull models
 - comparative-static analysis, 1000
 - eviction probability, 1000
 - landowners, 1000
 - squatter housing price, 1000
 - superfluous evictions, 1000
 - TVA. *See* Tennessee Valley Authority (TVA)
- U**
- Universities
 - county-industry panel, 1264–1265
 - deliberate policy, 1266
 - empirical analysis, 1264
 - evidence, 1265–1266, 1267–1268
 - factors, 1264
 - labor productivity, 1267
 - municipality-level panel, 1266
 - net innovation and productivity gains, 1267
 - types, 1265
 - Urban crime. *See* Crime
 - Urban form and homeownership
 - components, 1324
 - county regulations, 1323
 - density restrictions, 1325
 - homeownership rate, 1324–1325
 - housing supply regulation, 1322–1323
 - housing unit density, 1323–1324
 - land use policy, 1323
 - types, 1323
 - Urban growth boundaries (UGBs), 1446–1447
 - Urbanization
 - dynamics, 1415–1416
 - economic development, 1400
 - land-rights insecurity, 1400
 - livelihoods, 1400
 - slum dwellers, 1400
 - squatting, 1400
 - US Bureau of Economic Analysis (BEA), 1078
 - US context *vs.* Europe, 1088

- US Geological Survey (USGS), 1302
- US housing programs, 895–897

V

- Victimization, mental disorders, 1526
- Victims and would-be offenders interactions
 - blacks and Latinos, 1565
 - castle doctrine and stand-your-ground laws, 1562
- CompStat system, 1566
- deterrence and selection, 1555–1556
- incarceration/rehabilitation, 1552–1553
- lower rates per capita, 1552
- marginal and average hit rates, equality, 1563–1564
- Peer effects and strategic complementarity, 1559–1561
- police stops, 1562–1567
- private actions and displacement, 1553–1555
- racial and ethnic affiliation, 1562
- racial bias, 1564
- racial profiling, law enforcement agencies, 1562–1563
- racial residential segregation, 1557–1559
- “specific and articulable facts”, 1564
- statistical discrimination, 1565–1566
- stereotypes, 1556–1557
- street vice, homicide and effective law enforcement, 1553
- striking disparity across groups, 1563
- target and selection, 1552
- Vigdor’s model, 1087
- Violence reduction initiatives, 1546–1547
- Virtual office websites (VOWs), 856–857
- Voting, tax competition
 - equilibrium, municipalities, 1153
 - income distribution, 1153
 - Pareto distribution, 1155
 - population composition, 1154
 - sorting equilibrium, 1155

W

- Weak claim of criminogenic characteristics
 - brain functioning, structure, in utero experience and genetics, 1532–1534
 - causal effect, military service, 1534
 - description, 1528
 - education, 1529–1530
 - psychological and character traits, 1531–1532

- Weak claim of criminogenic characteristics
 - (*Continued*)
 - seafood consumption and homicides per capita, 1534
 - traits, 1528–1529
 - Welfare analysis
 - commuting technology, 1361–1362
 - counterfactual, 1376
 - infrastructure, 1361–1362
 - labor mobility, 1363
 - price index, 1361–1362
 - real income, 1362–1363
 - Welfare implications, regulation
 - effects, 1329
 - empirical efforts, 1328
 - land use controls, 1329
 - model-based assessment, 1327
 - population density, 1328
 - Wharton Residential Land Use Regulation Index (WRLURI), 1328, 1329
 - Welfare to Work Voucher evaluation, 897–899
 - Wharton Residential Land Use Regulation Index (WRLURI)
 - characterization, 1299
 - feature, 1301
 - and Gyourko–Saiz–Summers (GSS), 1299
 - metropolitan areas, 1299, 1300*t*
 - White-collar crimes, 1521–1522
 - WRLURI. *See* Wharton Residential Land Use Regulation Index (WRLURI)
- Z**
- ZFUs. *See* Zones Franches Urbaines (ZFUs)
 - Zipf's law, 1472–1473
 - Zipf's rank-size rule, 1141–1142
 - Zones Franches Urbaines (ZFUs)
 - enterprise zones model, 1246
 - heterogeneity, 1249
 - US programs, 1246
 - Zoning
 - Glaeser–Schuetz–Ward (GSW), 1301, 1302
 - and land values, 1304–1305
 - and public transit, 1303
 - regulations, 1293–1294