Essays in History and Spatial Economics with Big Data

Sun Kyoung Lee

Submitted in partial fulfillment of the
requirements for the degree
of Doctor of Philosophy
in the Graduate School of Arts and Sciences

COLUMBIA UNIVERSITY

2019

Abstract

Essays in History and Spatial Economics with Big Data

Sun Kyoung Lee

This dissertation contains three essays in History and Spatial Economics with Big Data. As a part of my dissertation, I develop a modern machine-learning based approach to connect large datasets. Merging several massive databases and matching the records within them presents challenges — some straightforward and others more complex. I employ artificial intelligence and machine learning technologies to link and then analyze massive amounts of historical US federal census, Department of Labor, and Bureau of Labor Statistics data.

The transformation of the US economy during this period was remarkable, from a rural economy at the beginning of the 19th century to an industrial nation by the end. More strikingly, after lagging behind the technological frontier for most of the nineteenth century, the United States entered the twenty-first century as the global technological leader and the richest nation in the world. Results from this dissertation reveal how people lived and how the business operated. It tells us the past that led us to where we are now in terms of people, geography, prices and wages, wealth, revenue, output, capital, numbers, and types of workers, urbanization, migration, and industrialization.

As a part of this endeavor, the first chapter studies how the benefits of improving urban mass transit infrastructures in cities are shared across workers with different skills. It exploits a unique historical setting to estimate the impact of urban transportation infrastructure: the introduction of mass-public transit infrastructure in the late nineteenth and twentieth-century New York City. I linked individual-level US census data to investigate how urban transit infrastructure differentially affects the welfare of workers with heterogenous skill.

My second chapter measures immigrations' role in US rise as an economic power. Especially, this chapter focuses on a potential mechanism by which immigrants might have spurred economic prosperity: the transfer of new knowledge. This is the first project to

use advances in quantitative spatial theory along with advanced big-data techniques to understand the contribution of immigrants to the process of U.S. economic growth. The key benefit of this approach is to link modern theory with massive amounts of microeconomic data about individual immigrants—their locations and occupations—to address questions that are extremely difficult to assess otherwise. Specifically, the dataset will help the researchers understand the extent to which the novel ideas and expertise immigrants brought to U.S. shores drove the nation's emergence as an industrial and technological powerhouse.

My third chapter exploits advances in data digitization and machine learning to study intergenerational mobility in the United States before World War II. Using machine learning techniques, I construct a massive database for multiple generations of fathers and sons. This allows us to identify “land of opportunities": locations and times in American history where kids had chances to move up in the income ladder. I find that intergenerational mobility elasticities were relatively stable during 1880-1940; there are regional disparities in terms of giving kids opportunities to move up, and; the geographic disparities of intergenerational mobility have evolved over time.

Contents

List of Figures vi

List of Tables ix

Chapter 1 Crabgrass Frontier Revisited in New York: Through the Lens of 21st century Data 1

1.1 Introduction: Crabgrass Frontier Revisited 4

1.2 Data and Methodology 7

1.2.1 New Population Data on Suburbanization in the US: 1870-1940 . . . 7

1.2.1.1 Neighborhood changes from repeated cross-sectional data . . 8

1.2.1.2 New longitudinal database and dynamic neighborhood changes 9

1.2.2 Record Linking 10

1.2.3 Geographic Information Harmonization 12

1.2.4 Transit Network 13

1.2.5 Geographic Definition 14

1.3 New York City Background: 1870-1940 17

1.3.1 Population Growth 17

1.3.2 Income and Occupation Trends in NYC 18

1.4 Testing the Specific Propositions of Crabgrass Frontier 19

1.4.1 Did Population Grow in More Suburban Areas? 19

1.4.2 Did the Rich Leave the Center of the City? 28

1.4.3 Were the People Who Moved Into the Periphery Richer than Original Residents of the Periphery? 50

1.4.4 Income Changes between the Core and the Periphery 72

1.4.4.1 Occupational income 73

1.4.4.2 Decomposition of the various flows among the core, the periphery, and the rest of the world 80

1.5 Conclusion 91

Chapter 2 European Immigrants and the United States' Rise to the Technological Frontier **93**

2.1 Introduction 95

2.2 A guiding example: The Story of Heinrich Engelhard Steinweg 98

2.3 Primary Data Sources and Methodology 100

2.3.1 Historical Data on Immigrants, Local Productivity and Innovation . . 101

2.3.1.1 *The Immigration Database: Immigration Records and Passenger Lists (1820 - 1914)* 101

2.3.1.2 *The Complete Count Federal Demographic Decennial Census (1850-1940)* 102

2.3.1.3 *Measuring Productivity: The Historical Census of Manufacturers (1870 - 1929)* 103

2.3.1.4 *Measuring Innovative Activity: Data on Patenting* 104

2.3.2 Dataset Construction: Record Matching And Harmonization 105

2.3.2.1 Record Matching: Linking Immigrants and the Complete Count Federal Demographic Decennial Census, 1850-1940 . 105

2.3.2.2 Record Matching: Linking US Patent Records and the Complete Count Federal Demographic Decennial Census, 1850-1940107

2.3.2.3 Data Harmonization and Integration at the Country-Industry Level . . . 107
2.4 Setup . . . 108
2.4.1 Firms and Innovation . . . 109
2.4.2 Labor Supply . . . 112
2.4.3 Aggregation . . . 114
2.4.4 Dynamical Equilibrium . . . 116
2.4.5 Innovation in Space and Time . . . 117
2.5 The Spatial Balanced Growth Path . . . 118
2.6 Reduced form evidence . . . 120
2.6.1 Constructing an instrument for the allocation of immigrants . . . 120
2.6.2 Immigrants and Regional Productivity Growth . . . 121
2.6.3 Immigrants and Regional Patent Activity . . . 125
2.6.4 Immigrants and Patent Novelty . . . 127
2.7 Counterfactuals . . . 127
2.8 Conclusions . . . 130
Chapter 3 American Intergenerational Mobility in History and Space **131**
3.1 Introduction . . . 133
3.2 Record Linking Methodology and Imputation of Incomes . . . 135
3.2.1 Record Matching Overview . . . 135
3.2.1.1 Traditional Iterative Decision Tree Method . . . 136
3.2.1.2 Automated Probabilistic (EM) Algorithm . . . 137
3.2.1.3 Machine Learning Method (This Paper) . . . 137
3.2.2 This Paper: A (Supervised Discriminative) Machine Learning Approach for Record Linking . . . 137
3.2.3 Methodology to Construct Position in the Income Distribution . . . 139
3.2.4 Final Sample Selection and Summary Statistics . . . 142

3.3 Results 144

3.4 Historical Intergenerational Mobility in Space 148

3.5 Conclusion 149

Bibliography **153**

Appendix A for Chapter 1 **157**

A.1 Methodology 158

A.1.1 Machine Learning Approach of Record Matching 158

A.1.2 Record Linking in Practice: Innovations 159

A.1.2.1 Female Record Linking 159

A.2 The Transportation Revolution 160

A.2.1 Intra-city transit infrastructure changes 160

A.2.2 Inter-city transit infrastructure changes 164

A.3 Supplementary Figures 167

A.3.1 Different Land Use Creation 167

A.3.1.1 Residential Land Use Construction 167

A.3.1.2 Commercial Land Use Construction 169

A.4 Theoretical framework 170

A.4.1 Setup 170

A.4.2 Migration 171

A.4.2.1 Migration decision on which labor market to face 171

A.4.2.2 Neighborhood decision and commuting costs 172

A.4.3 Equilibrium 173

Appendix B for Chapter 2 **175**

B.1 Proofs 176

B.1.1 Proof of Proposition 2.4.1 176

B.1.2 Deriving the labor supply relationships 177

B.1.3 Proof of Lemma 2.4.3 178

B.1.4 Characterization of the Balanced Growth Path (BGP) 179

List of Figures

11 Geographic Boundary Definition 16

12 Population Trend Over Time by Borough 17

13 Mean Occupational Income Trend: Men and Women 19

14 Population Density Result Coefficients and Confidence Intervals (with zeros) . . 22

15 Population Density Result Coefficients and Confidence Intervals (without zeros) 23

16 ln(*Population Density*) Against Distance From Center 24

ln(*Population Density*) Against Distance From Center 25

ln(*Population Density*) Against Distance From Center 26

17 Population Density 27

18 % Change of Population Density 28

19 Neighborhood-level Mean Income Differences between Leavers and Stayers, 1880-1900 31

110 Neighborhood-level Mean Income Differences between Leavers and Stayers, 1900-1910 35

111 Neighborhood-level Mean Income Differences between Leavers and Stayers, 1910-1920 39

112 Neighborhood-level Mean Income Differences between Leavers and Stayers, 1920-1930 43

113 Neighborhood-level Mean Income Differences between Leavers and Stayers, 1930-1940 47

List of Figures Cont'd

114 Neighborhood-level Mean Income Differences between Entrants and Stayers, 1880-1900 53

115 Neighborhood-level Mean Income Differences between Entrants and Stayers, 1900-1910 57

116 Neighborhood-level Mean Income Differences between Entrants and Stayers, 1910-1920 61

117 Neighborhood-level Mean Income Differences between Entrants and Stayers, 1920-1930 65

118 Neighborhood-level Mean Income Differences between Entrants and Stayers, 1930-1940 69

119 Mean Occupationa Income 74-77

120 % Change of Mean Occupational Income 78

% Change of Mean Occupational Income 79

% Change of Mean Occupational Income 80

121 Mean Income and Magnitude of Flows Across Migration-type: 1880-1900 85

Mean Income and Magnitude of Flows Across Migration-type: 1900-1910 86

Mean Income and Magnitude of Flows Across Migration-type: 1910-1920 87

121 Mean Income and Magnitude of Flows Across Migration-type: 1920-1930 88

Mean Income and Magnitude of Flows Across Migration-type: 1930-1940 89

122 The Core and Periphery Stayers' Income and Ratio 90

21 Heinrich Steinweg in the *Hamburg Passenger Lists, 1850-1934* 98

22 Henry Steinway in the *US Census Schedules 1860 and 1870* 99

23 Mr Steinway's Pianoforte patent and Steinway & Sons in the *US Census of Manufacturers 1880* 100

List of Figures Cont'd

24 Data Construction: Combining the Data Sets 101

25 Constructing the Instrument 121

26 Immigrants and Productivity Growth 124

27 Immigrants and Patent Activity 127

31 Intergenerational Mobility Coefficients: Son's Average Income Position 145

32 Intergenerational Mobility Coefficients: Son's Probability of Being in the Top and Bottomn Quartile. 147

33 Intergenerational Mobility Coefficients Across States: Son's Average Income Position 149

34 Intergenerational Mobility Coefficients Across SEAs: Son's Average Income Position150

35 Intergenerational Mobility Coefficients Across SEAs: Predicted Percentile of Son with Poor Father 151

A1 Total number of intra-city railway, subway stations 161

A2 Evolution of Spatial Links by the Elevated Railway, Subways 162

A3 The Els, Subways-Based Transit Access Measures by Decade 163

A4 Evolution of Inter-City Transit Infrastructure by Construction Year 165

A5 Evolution of Spatial Links by Bridges, Ferry, Tunnel 166

A6 New Construction of Residential-Use Land 167

New Construction of Residential-Use Land 168

A7 New Construction of Commercial-Use Land 169

New Construction of Commercial-Use Land 170

List of Tables

11 Population Density (with zeros) 21

12 Population Density (without zeros) 22

13 Logit Results between Leavers and Stayers at the City level: 1880-1900 Males . 32

14 Multinomial Logit Results between Leavers and Stayers at Neighborhood Level: 1880-1900 males 33

Multinomial Logit Results between Leavers and Stayers at Neighborhood Level: 1880-1900 males (cont.) 34

15 Logit Results between Leavers and Stayers at the City level: 1900-1910 males . 36

16 Multinomial Logit Results between Leavers and Stayers at Neighborhood Level: 1900-1910 males 37

Multinomial Logit Results between Leavers and Stayers at Neighborhood Level: 1900-1910 males (cont.) 38

17 Logit Results between Leavers and Stayers at the City level: 1910-1920 males . 40

18 Multinomial Logit Results between Leavers and Stayers at Neighborhood Level: 1910-1920 males 41

Multinomial Logit Results between Leavers and Stayers at Neighborhood Level: 1910-1920 males (cont.) 42

19 Logit Results between Leavers and Stayers at the City level: 1920-1930 males . 44

110 Multinomial Logit Results between Leavers and Stayers at Neighborhood Level: 1920-1930 males 45,46

List of Tables Cont'd

111 Logit Results between Leavers and Stayers at the City level: 1930-1940 males . 48

112 Multinomial Logit Results between Leavers and Stayers at Neighborhood Level: 1930-1940 males 49, 50

113 Logit Results between Entrants and Stayers at the City level: 1880-1900 males 54

114 Multinomial Logit Results between Entrants and Stayers at Neighborhood Level: 1880-1900 males 55, 56

115 Logit Results between Entrants and Stayers: 1900-1910 males 58

116 Multinomial Logit Results between Entrants and Stayers at Neighborhood Level: 1900-1910 males 59, 60

117 Logit Results between Entrants and Stayers at the City level: 1910-1920 males . 62

118 Multinomial Logit Results between Entrants and Stayers at Neighborhood Level: 1910-1920 males 63, 64

119 Logit Results between Entrants and Stayers at the City level: 1920-1930 males . 66

120 Multinomial Logit Results between Entrants and Stayers at Neighborhood Level: 1920-1930 males 67, 68

121 Logit Results between Entrants and Stayers at the City level: 1930-1940 males . 70

122 Multinomial Logit Results between Entrants and Stayers at Neighborhood Level: 1930-1940 males 71,72

List of Tables Cont'd

21 Linking Immigration Records and the US Population Census 106

22 Immigrants and Productivity Growth (Equation (2.31)) 123

23 Immigrants and Patent Activity 126

24 Immigrants and Patent Novelty 128

31 Model Fit: Manufacturing Employment - Alternative Specifications 140

32 Income Extrapolation: Predictive Power of Income Extrapolation based on 1940 and 1950 income in and out of sample. 141

33 Population Census Samples: Alternative Specifications 142

34 Twenty Years Matched Sample Size: Alternative Specifications 143

This page intentionally left blank

Acknowledgements

Throughout my doctoral studies, I have been indebted than I could acknowledge here.

I would like to thank my sponsor Donald R. Davis. I chose to pursue Economics Ph.D. at Columbia University with my keen interests in international trade and policy, with the hope that I study how international trade policies may affect the welfare of developing countries. However, upon arriving at Columbia and taking my field courses with Don, my geographic scope of interests gradually switched from "countries" to "cities and neighborhoods." Don's excitement about urban issues was simply transformative and powerful and I will be always grateful of him to guide me through the academic journey at Columbia. I have interacted with Don as his protégé, Teaching Assistant, and Research Assistant for years. Don's philosophy, exactness and charisma are truly inspirational and I am very grateful of Don for guiding me through the long journey.

I have been extremely lucky to have Dan O'Flaherty as my advisor. If I could call Don as my academic father, Dan has been my academic mother. Through his finesse, Dan has taught me right from wrong, dos and don'ts; Dan has given me courage when I was in self-doubt, showed me how to achieve things through small steps. Most importantly, Dan has taught me to stay self-critical, stay respectful for everyone and everything around us.

Suresh Naidu has been an invaluable advisor, and bouncing ideas with Suresh has felt like oxygen to the thought process. If Don and Dan are my roots for pursuing Urban Economics, Suresh is the soil for my interests in Economic History and Big Data through computer science techniques. When I was stuck in major steps, talking to Suresh about the problems gave me a very fresh perspective of how to approach the problems.

I have greatly enjoyed collaborating with Costas Arkolakis at Yale University. I have begun the endeavors of working on Big Data Infrastructure project with Costas since 2016. Costas's bulldozer-like spirit added a big impetus to the project. At times, I was overwhelmed by the scope and amount of works that I had to deal with on such big-scale projects. However, I cannot thank enough Costas for believing in me and assuring me that I would pull it off in the end through (borrowing his expression) *"working crazy hard."* Costas's faith in my talent has kept me going through difficult and arduous phases that required me a great deal of patience. As one of my advisors said, "*a beautiful complementarity between his work and your work*" has created massive synergy and I am very grateful for Costas's extraordinary patience and firm beliefs in me.

While collaborating with Costas, I began working with Michael Peters at Yale and I must say that working with Michael has been such a rewarding experience. "Learning-by-doing" has been truly real while collaborating with Costas and Michael. By closely working together and witnessing how they approach every step of the problem, I have developed my way of solving the problem and comparative skills. Thanks to our collaboration, three of us have won prestigious grants from the National Science Foundation and the National Bureau of Economic Research.

I have benefitted tremendously from being surrounded by extremely talented economists at Columbia and I would like to thank them. Especially, Professor José Scheinkman has been simply a marvelous figure. José has provided me invaluable insights especially on the big picture of the projects that I am pursuing. José has been incredibly approachable and I deeply regret not talking to him much earlier! I also would like to thank Réka Juh*á*sz for being a great mentor and showing me ropes of achieving things by staying focused in the present.

My dear friends are have been wonderful company in my long, academic journey. Especially, I would like to thank Colin Hottman, Keshav Dogra for their warm encouragement and helped me get through difficult times; I also would like to thank Jing Zhou, Anurag Singh and Haaris Mateen for their genuine and sincere friendship. Our friendship has grown over years-long over take-out meals and late-night studies together, and that sense of solidarity was deeply comforting to my soul. Finally, I want to thank my best friends Jiwon Seo, and Catheirne Junghyun Seo for our ever-growing friendship. The Seo sisters' beautiful souls helped me land on my feet; it was their love that made me feel beautiful inside.

Beyond the academic environment, I want to thank my family for their endless support and love. My father, an urban planner, would discuss with me over the dinner table how politics can affect the infrastructure budget and how the survival of flea markets is vital to the neighborhood community and my interest in urban policy grew organically. My mother's support and kindness have been simply exemplary and I cannot imagine the counterfactual world without her warmth. I also want to thank my sister for being the rock as my closest friend and sending me numerous, supportive hand-written letters that brought so much strength. My sincere gratitude for my brother-in-law, Dong-Joon Lee, for his thoughtful and generous support including sending me very carefully selected snacks from Korea so that I can feel at home even for a while. My nephew Maru has been extremely understanding of my very long absence and I would like to thank him again for his unique and adorable way of expressing his love to me. My furry friend, Artemis, has brought me joy and laughter every night and I am very grateful for her spirit and her sacrifice of staying up endless nights for me to soldier on.

Finally, I would like to dedicate my dissertation to my grandfather In-Young Lee who is no longer with us. Words will be never enough to describe how much I respect and love him. He has been the biggest supporter of my education, shared with me his mantra "`girls, be ambitious and empower yourself through education`" as long as I could remember. I am extremely sorry that I could not spend the very last moment of his life as I was a thousand miles away with my studies. The least I could do is to dedicate my dissertation to the man who has supported me and my education with all of his heart.

To my parents, and my grandfather In-Young Lee

Chapter 1

Crabgrass Frontier Revisited in New York: Through the Lens of 21st-century Data

Sun Kyoung Lee[1]

[1] I thank Don Davis, Suresh Naidu, Dan O'Flaherty for their continued invaluable advice and support. Thanks also to Costas Arkolakis, Jason Barr, Devin Bunten, Ingrid Gould Ellen, Réka Juhász, Michael Peters, Stephen Redding, José Scheinkman, Anurag Singh, David Weinstein, and seminar and colloquium participants at Columbia, and NYU Furman Center. I thank Jean Roth, Dan Feenberg, Mohan Ramanujan at National Bureau of Economic Research for their help with historical US Federal Census data, IPUMS team at Minnesota Population Center, Allison Shertzer for providing access to Historical GIS data, Mishka Vance at New York Public Library and Helley Choiniere and Deseree Alden at New York Transit Museum Archive for providing me various resources. I am grateful to `Smartystreets.com` for providing unlimited address validation service. Financial support from National Science Foundation (under grant \#1658827, grant \#1831524), Economic History Association, Lincoln Institute, Program for Economic Research at Columbia University, and Kwan Jeong Educational Foundation is gratefully acknowledged. All errors are my own.

Crabgrass Frontier Revisited in New York :

Through the Lens of 21st-century Data

Sun Kyoung Lee

Columbia University

Abstract

Jackson's famous *Crabgrass Frontier: The Suburbanization of the United States* (1985) argues that when American cities suburbanized in the early nineteenth century, the richest households moved from the core to the periphery, the poorest stayed in the core, and the households that moved to the periphery were richer than those who were there before them. I study the gradual process of prewar suburbanization in America's biggest city, New York City, between 1870 and 1940. During this time there were huge transportation infrastructure improvements at both intra- and inter-city level, and there was gradual suburbanization, just as in Jackson (1985). I construct a historical longitudinal database that follows individuals to analyze how the migration patterns differ across workers with different income (skills). Rich people on average did not leave the core and poor people on average did not stay. New suburbanites to the city periphery were not richer than the people who already lived at the periphery. Jackson's fundamental claim about the growth of high income at the edge relative to the center still holds true for my study period. However, I show the mechanism behind this change and show that this relative change in income growth at the edge

did not result from a simple shuffling of rich and poor. Up until the Great Depression, flows of migrants from and to outside the metropolitan area were the dominant force in changing average income. Richer people from outside NYC metropolitan area migrated to the periphery and poorer people from outside NYC metropolitan migrated to the core. The people from the city core who left the metropolitan area were far richer than the people from the periphery who left the metropolitan area. Furthermore, people who stayed at the periphery got richer as the metropolis grew. Many readers have interpreted *Crabgrass Frontier* as the story of America's suburbanization always and everywhere, and so my finding that two of the major propositions in that book and the mechanism behind income growth at the edge do not apply to 1870-1940 New York has implications beyond local history.

1.1 Introduction: Crabgrass Frontier Revisited

> *"Our property seems to me the most beautiful in the world. It is so close to Babylon that we enjoy all the advantages of the city, and yet when we come home we are away from all the noise and dust."*[1]

Jackson (1985)'s *Crabgrass Frontier: The Suburbanization of the United States* remains one of the most influential books ever written on urban history and on American cities. One of the main ideas in the book is that the rich began the flight from the city first — something that the middle classes eventually emulated as city tax rates skyrocketed and those on the lower end of the economic stratum moved into the city.

Jackson (1985) found that suburbanization, a phenomenon that started no later than the early 19th century, was accompanied by enormous growth in metropolitan size and rapid population growth on the periphery, an absolute loss of population at the center, and an increase in the average journey to work, and a rise in the socioeconomic status of suburban residents.

However, Jackson did not have the benefit of the datasets and quantitative analysis techniques that we have now. In particular, he could not follow individuals. He could observe, for instance, that suburbs gained population and that the central parts of cities lost population, but he did not know whether any individual moved from the city to the suburb. With only periodic snapshots of aggregates and no guarantee that anyone in any snapshot is in any other, we cannot begin to think how events affected individuals. While formal welfare analysis is beyond the scope of the current paper, the longitudinal analysis that I perform here is a necessary prologue to any such work.

My study period (1870-1940) occurs after Jackson's study (i.e. 1815-1875) and before the major introduction of highways in the US (Baum-Snow (2007)). I concentrate on New

[1] Jackson (1985) cites the letter in cuneiform on a clay tablet, which was a letter to the King of Persia in 539 B.C.. Jackson (1985) argues Boston, Philadelphia, and New York established suburbs well before the Revolutionary War, and this letter represents the first extant expression of the suburban ideal.

York City. I investigate three of the major conclusions of *Crabgrass Frontier:*

1. That the relative population of more suburban areas increased

2. That the richest people were the ones who led the movement from the center of the city to the periphery, and poor people stayed in the center of the city

3. That the people who moved from the center to the periphery were richer than the people living in the periphery

I investigate each of these propositions in turn. Notice that second and third propositions are explicitly longitudinal statements that repeated cross-section data cannot examine.

To bring Jackson's work to the 21st century, I create a micro longitudinal database of individuals by linking US demographic census records. These new datasets provide a very high level of geographic resolution and help shed light on the evolution of neighborhoods over a long time horizon as transportation infrastructure was developed. I link individual-level US demographic census records from 1870 to 1940 (every decade, except for 1890 as the 1890 population census was lost due to fire), to track individuals' residential locations in relation to transit infrastructure-driven transit access change. I decompose how neighborhood changes were driven by out-migration and in-migration of individuals with different socioeconomic characteristics, along with the incumbents' income increase when intra-city transit infrastructure improved market access.

Through the above approach, I "let the data speak" about the process of suburbanization in the biggest city in America. Regarding Jackson's three points, I find

1. Yes, population decentralized

2. No, the people who stayed in the center of the city were richer than the ones who left the city center

3. No, the people who moved to the periphery were poorer than the people already living there

4. Relatedly, richer people from outside NYC metropolitan area migrated to the periphery and poorer people from outside NYC metropolitan area migrated to the core; the people from the city core who left the NYC metropolitan area were far richer than the people from the periphery who left the metropolitan area; furthermore, people who stayed at the periphery got richer as the metropolis grew.

Jackson's fundamental claim about the growth of high income at the edge relative to the center still holds true for my study period. However, I show the mechanism behind this change and show that this relative change in income growth at the edge did not result from a simple shuffling of rich and poor.

Up until the Great Depression, flows of migrants from and to outside the metropolitan area were the dominant force in changing average income. Richer people from outside NYC metropolitan area migrated to the periphery and poorer people from outside NYC metropolitan migrated to the core. The people from the city core who left the metropolitan area were far richer than the people from the periphery who left the metropolitan area. Furthermore, people who stayed at the periphery got richer as the metropolis grew.

To be sure, I am studying only one city for one period, and it is a period outside Jackson's explicit study period. But the city is America's largest, and the period encompasses New York's greatest growth and most dramatic change. Many readers have interpreted *Crabgrass Frontier* as the story of America's suburbanization always and everywhere, and so my finding that two of the major propositions in that book and the mechanism behind income growth at the edge do not apply to 1870-1940 New York has implications beyond local history.

Several research projects explain the central city population decline. For example, Baum-Snow (2007) demonstrates that the construction of new limited-access highways caused central city population decline. Boustan (2010) focuses on sorting where white households left central cities due to racial preferences. Relative to the aforementioned papers, I use a panel data of individuals that enables me to decompose the relative magnitudes of the flows among entrants, leavers and stayers and its associated income differences.

This paper also relates to the large reduced-form empirical literature on transport infrastructure, including Banerjee et al. (2012), Baum-Snow (2007), Donaldson (2010), Donaldson and Hornbeck (2013), Faber (2013), Duranton et al. (2013), Michaels (2008). This paper also contributes to the literature on the internal structure of the city, through a quantitative analysis of economic geography. While there has been extensive development of economic geography in the past few decades (Fujita and Ogawa (1982), and Lucas and Rossi-Hansberg (2002)), there is growing empirical literature. Especially, the structural estimation approach has been implemented in studying the allocation of economic activity, including Ahlfeldt et al. (2012), Allen and Arkolakis (2013), Allen et al. (2015), Heblich et al. (2018), Monte et al. (2015), and Tsivanidis (2018). Especially, Heblich et al. (2018) use the invention of steam railways in the 19th century London to document the role of separating the workplace and residence in supporting concentrations of economic activity. Tsivanidis (2018) evaluates the effect of the world's largest Bus Rapid Transit in Bogota, Colombia and show the gains of improving transit in cities may differ across skill groups.

The remainder of the paper is structured as follows. Section 2 discusses the data and methodology. Section 3 discusses the relevant background of the study. Section 4 discusses the findings of revisiting some propositions of *Crabgrass Frontier* in New York during the study period. Finally, Section 5 concludes.

1.2 Data and Methodology

1.2.1 New Population Data on Suburbanization in the US: 1870-1940

I use restricted-access complete count individual-level US Federal Demographic Census records from 1870 to 1940 to analyze skill-based internal migration in relation to transit infrastructure. These individual-level census records provide rich socioeconomic and demographic information such as occupation, industry, race, and family characteristics along with the residential location. However, complete-count population censuses only exist in cross-sectional format and they do not have a time-invariant individual identifier(s). As following the same individuals over time is essential, I use a "machine learning" approach to follow the same individuals over time. I summarize the record linking criteria and procedure in Subsection 1.2.2, and details on individual-level record linking are available in the Appendix.

1.2.1.1 Neighborhood changes from repeated cross-sectional data

In this paper, I use the 1950 Census Bureau occupational classification system (henceforth, OCC1950)-based occupational measures of income and education to enhance comparability across the years. Ruggles et al. (2019) coded occupation-based values according to the 1950 classification. Throughout the analysis, I use OCC1950-based occupational income score (variable called "OCCSCORE") as measures of occupational standing. OCCSCORE is a constructed 2-digit numeric variable that assigns occupational income scores to each occupation in all years of pre-1950 US census which represents the median total income (in hundred of 1950 dollars) of all persons with that particular occupation in 1950.[2]

This approach of using OCC1950-based OCCSCORE controls for inflation and is widely used in the literature to measure individuals' skills. OCC1950 is divided into 10 social classes and 269 occupational groups and has been the US standard for occupational coding due to its strength in comparability across years. However, it has potential shortcomings of not reflecting the relative wage changes, and relative wages may be different across locations. Despite these potential shortcomings, this approach allows me to document neighborhood changes in terms of residents' skills over time (the US Federal demographic census records asked neither one's income nor educational attainment until 1940).

Regarding sources of neighborhood changes, for neighborhoods to change in terms of composition of residents, at least one of three things must hold true (Ellen and O'Regan

[2] Detailed description of "OCCSCORE" and "OCC1950" are available here: https://usa.ipums.org/usa-action/variables/OCCSCORE#codes_section, https://usa.ipums.org/usa-action/variables/OCC1950#description_section

(2011)) — 1. new entrants to the neighborhood must have different socioeconomic characteristics than the neighborhood average **(selective entry)**; 2. households exiting the neighborhood must have different socioeconomic characteristics than the average (**selective exit**); 3. those remaining in the neighborhood must experience the socioeconomic changes (**incumbent changes**). I can follow all three groups as I seek to match every individual appearing in the US Demographic census records from 1870 to 1940 (every decade, except for 1890 as original 1890 Census records were lost due to fire). The above approach of following all three groups over time requires a longitudinal individual-database and in the following section, I provide more detail about record-linking process.

1.2.1.2 New longitudinal database and dynamic neighborhood changes

I analyze the longitudinal data of individuals and document how different income groups migrated differently.[3] The longitudinal tracking of individuals is essential to revisit the *Crabgrass Frontier* propositions for the following reason: suppose one observes a city or neighborhood at two different times, one can observe only how the aggregates changed. Given any sequence of aggregates, there are a huge number of different individual sequences that can produce them, and those different collections of individual sequences have different welfare interpretations. For instance, if one observes only that average income in a city rises between 1870 and 1880, it is unclear whether the people who lived in that city in 1870 stayed there and prospered, or the people who lived there in 1870 suffered and fled the city only to be replaced by richer people who were also losing income but from a higher starting point.

Jackson (1985) argues that when transit infrastructure improved, the rich left the older areas, whereas the poor stayed in the older areas. Therefore, in order for me to revisit these

[3] I create longitudinal database by linking individual demographic census records for both males and females. For female records, however, due to last name change traditions during the study period, I use the marital status information of female records at between two census periods and link only females where the marital status had *not* changed (i.e. single in both periods, married in both periods, or some other cases where last name changes do not typically happen such as married in earlier period and widowed in the later period). For 1870 census records, as the marital status information is not available, I did not link female records between 1870 and 1880.

propositions, I need longitudinal data of individuals with different skills (or incomes). To do this, I follow everyone in the US census records (not a sample) during the study period including people who entered, people who left, and people who stayed in neighborhoods in the city between two adjacent censuses. I classify them into "entrants", "leavers" and "stayers" based on their residential location-based migration-status at the neighborhood level. For every neighborhood in the city, "entrants" denote people who lived somewhere other than the particular neighborhood in the earlier period and then migrate into the particular neighborhood in the later period. "Stayers" denote the group of people who live in the same particular neighborhood in the city, whereas "leavers" denote the ones who lived in that particular neighborhood in the earlier period, and no longer live in that neighborhood in the later period. Details of the census record linking are available in the Appendix.

1.2.2 Record Linking

I implement a supervised discriminative machine learning approach to link historical records without time-invariant individual identifier(s). The essence of this approach is that I use training data (as "teaching-material") to train the algorithm on how to identify the potential matches based on certain discrepancies in the data.[4] I exploit the complete transcription of decennial federal census records from 1870 to 1940 except for 1890 (which was lost due to fire), and create a linked-individual longitudinal database across different census years.

Similar efforts of linking records using machine learning methods have been made by Goeken et al. (2011) who built the IPUMS linked individual samples using 1% samples of the 1850 to 1930 US population censuses and 1880 complete count census, and Feigenbaum (2015) who linked historical records of children in the 1915 Iowa State Census to their adult-selves in the 1940 Federal Census. Relative to the mentioned work, this project is far

[4] For example, Heinrich Engelhard Steinweg, the founder of prominent piano manufacturing company, *Steinway & Sons*, anglicized his names into "Henry E. Steinway." Therefore, in linking his records across censuses, string comparison measures called Jaro-Winkler distance of his first (Heinrich vs. Henry), middle (Engelhard vs E.) and last name (Steinweg vs Steinway) would show name discrepancies) even if his birth year and birthplace may be the same across different records

more extensive in the scope of matching as it links complete-count census records of the study period (i.e. 1870 to 1940). I create a training dataset which contain both "true" and "false" matches and their characteristics (e.g. some observations with "true" as an outcome would have same/very similar characteristics in terms of age, first and last name, parents' and his/her birthplaces whereas observations with "false" as an outcome would have quite different characteristics in terms of the above mentioned characteristics). In this case, the outcome is whether the matched records are "true" or "false" match, given the observed characteristics. By taking this training data, I build a prediction model, or learner, which will enable us to predict the outcome for new, unseen objects. A well-designed learner armed with a solid training dataset should accurately predict outcomes for new unseen objects.

I implement a supervised learning problem in the sense that the presence of outcome variable ("true" or "false" links) guides the learning process—in other words, the end-goal is to use the inputs to predict the output values. To summarize this process, I extract subsets of possible matches for each record and create training data in order to tune a matching algorithm so that the matching algorithm matches individual records by minimizing both false positives and false negatives while reflecting inherent noises in historical records. I explored various models for model selection, and I ultimately chose the random forest classification as it is *more conservative* in matching records and the number of unique matches are significantly higher than the standard Support Vector Machine model.

Also, I develop a record linking algorithm and methodology that links women's census records over time. Linking women's records is very rare because women's last name changed upon marriage in this period. Also relative to other traditional record linking methods where potential non-unique candidate matches are eliminated, I implement various ways to save more matches by including time-invariant family information.

1.2.3 Geographic Information Harmonization

The primary geographic units of the analysis are "Neighborhood Tabulation Areas" (hereafter, NTAs), each with at least 15,000 people in 2010 (there are 195 NTAs (neighborhoods) within the city). As datasets used in the analyses have different spatial units and/or the boundaries of the spatial unit constantly change, I create spatial crosswalks from historical spatial locations from various data sources (e.g. "enumeration district" in US census records) to NTAs so that NTAs can be a time-invariant, consistent geographic unit of analyses, and all datasets used in the analyses are harmonized and geolinked to NTAs.

An Enumeration District is a historical version of "census tract" where the historical US census enumerators recorded as administrative division smaller than counties (and wards which were extensively used in existing literature). As individual-level US Federal Demographic Census provides ED number, I can now aggregate the individual-level information to the neighborhood or similar geographic levels within the city. As GIS software enables researchers to know where these geographic units are in space, historical GIS effort of georeferencing ED images from microfilms and creating Geographic Information System (hereafter, GIS)-compatible shapefiles must be made to execute the analyses during the study period (i.e. 1870 to 1940).

This digitization effort has benefited from existing projects called the Urban Transition NHGIS (Logan et al, 2011) and Shertzer et al. (2016). I complement the existing sources by pushing time horizon and geographic scope—1880 Enumeration District boundary files of Manhattan and Brooklyn were obtained from the Urban Transition Historical GIS project; Shertzer et al. (2016) shared with me Manhattan and Brooklyn ED boundary files from 1900 to 1930. However, as Shertzer et al. (2016) mainly focus on studying the ten US largest cities, they did not digitize the relatively unpopulated areas of the Bronx, Queens, and Richmond. Therefore, I use the microfilm scan images of New York City Enumeration District maps of 1880-1940 and created historical GIS files for the remaining regions across time. For boroughs that microfilm scan images were not available in each period such as Queens county in 1900, Richmond County and Bronx county in 1910, I use detailed street and building information of residential addresses from the individual-level census records to locate which ED corresponds to each neighborhood. Stephen P. Morse's website has resources for ED finding tools for 1900 to 1940 censuses https://stevemorse.org/census/unified.html, and I mainly reference this website to check the conversion between different census years, and old street names and ED boundaries.

A major difficulty in making use of ED-level analysis using the above-mentioned boundary files is that the ED boundaries change considerably across time, making it extremely challenging to form consistent neighborhoods. Shertzer et al. (2016), for example, approach this problem by harmonizing ED data to temporally invariant geographically defined areas that they treat as "synthetic neighborhoods" to study neighborhood change. I approach this problem by taking the Neighborhood Tabulation Areas (called "NTAs") created by Department of City Planning in New York City.[5] I use ED shapefiles to create spatial crosswalks from ED boundaries to NTA neighborhoods over the study period. For every ED and every NTA, I aggregate the variables by aggregating the complete-count US Demographic census. Examples of such are total population, age, mean family size, occupation-based earning and education measures, marital status, and race.

1.2.4 Transit Network

I have collected various subway and elevated railway datasets, including the data on each station in the existing New York transit system. The year each station has opened was determined to estimate the subway opening, network, and station effects. Based on the compiled dataset, and evolution of subway and the elevated train network every decade (1870 to 1940) is documented. I use this transit network evolution, and I classify each neighborhood in the city as "transit hubs" as the core and "transit spokes" as the periphery of the city. "Transit hubs" are locations where transit infrastructures are extremely concentrated such

[5] Description and related GIS software-compatible files of Neighborhood Tabulation Areas is available here: https://www1.nyc.gov/site/planning/data-maps/open-data/dwn-nynta.page

as Downtown Manhattan and Midtown Manhattan, whereas “transit spokes” are locations where transit connections exit with low density but connected to transit hubs.

1.2.5 Geographic Definition

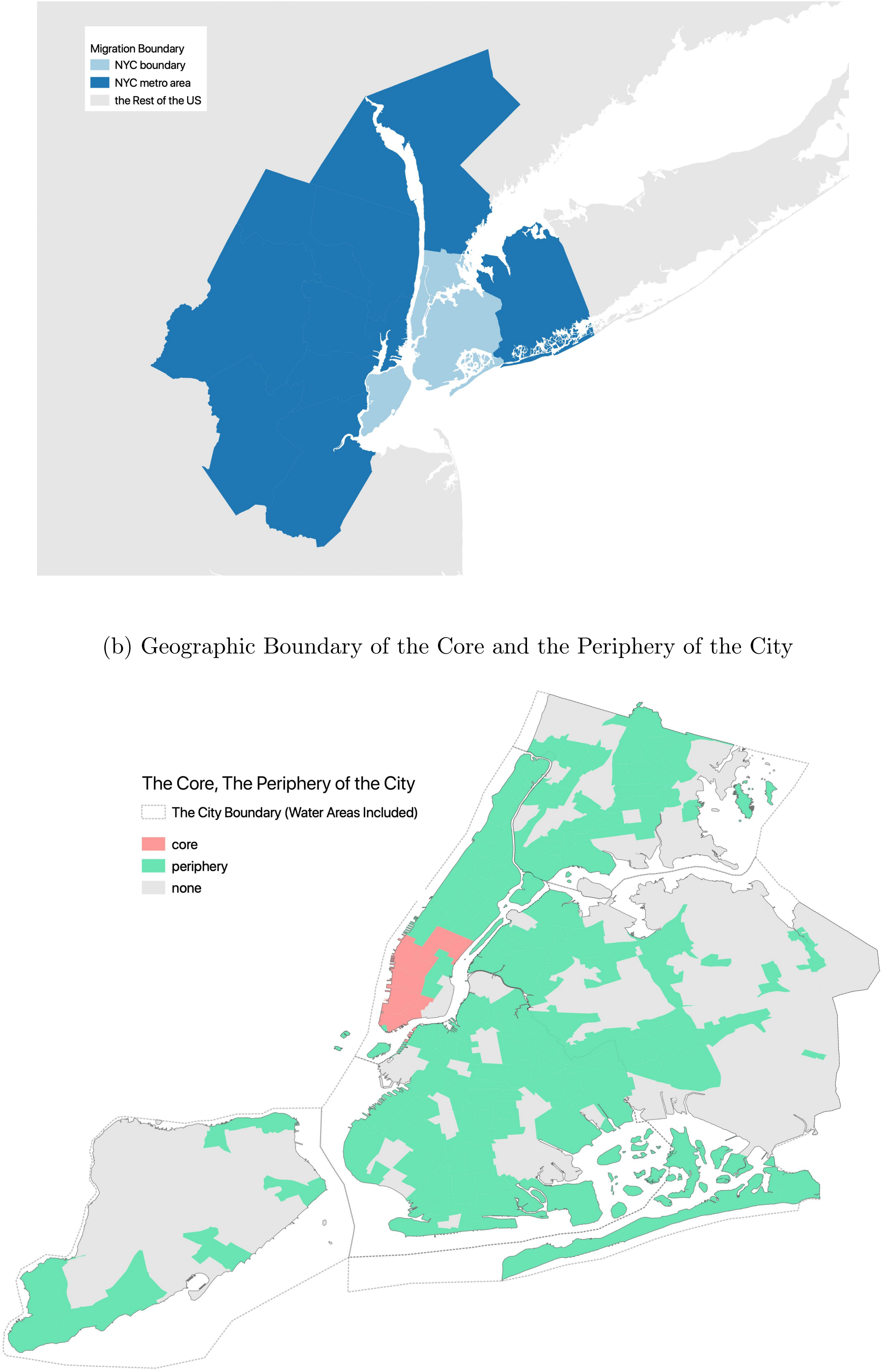
There are essential boundary definitions in Section 1.4.2, 1.4.3, and the decomposition analyses in Section 1.4.4. Here is the definition of the metropolitan area (or metro area), and how I define the core and the periphery of the city. Figure 11a shows geographic boundary of NYC, NYC metro area, and the rest of the country. A metro area, or metropolitan area, is a region consisting of a large urban core together with surrounding communities that have a high degree of economic and social integration with the urban core and I follow the IPUMS-definition, and delineation of the metro area of New York City.[6] IPUMS-delineation of the metro area of the city applies the 1950 Office of Management and Budget standards to historical statistics. This approach yields time-varying delineations of regions with high degree of economic and social integration with the urban core which is ideal for my study (i.e. Suffolk County, New York was not part of NYC metro area till 1920, however, as the economic integration between Suffolk county and NYC increased, Suffolk county became a part of NYC metro area since 1930). As in Figure 11a, I define 5 boroughs of New York City as the city (in Light blue), NYC metro area (in Dark blue), and the rest of the United States (in Light gray) by following IPUMS delineations of NYC metro areas.

Figure 11b shows the core and the periphery of the city. I define the core and the periphery neighborhoods in the city based on the intra- and inter-railway transit network over time (as in Section 1.2.4), and therefore the delineation of what makes up the core (in Pink) and the periphery (in Emerald green) of the city changes over time depending on the transit infrastructure at that time. The city is the union of core, periphery and the rest — transit hubs are the core of the city where the transit infrastructure is extremely well connected, whereas transit spokes make up the periphery of the city with low intensity of

[6] Description of a metropolitan area and definition is available here: https://usa.ipums.org/usa-action/variables/METAREA#description_section

transit network, and the rest of the city (in Light gray) are areas with no direct transit access.

Figure 11: Geographic Boundary Definition



(a) Geographic Boundary of NYC Metro Area

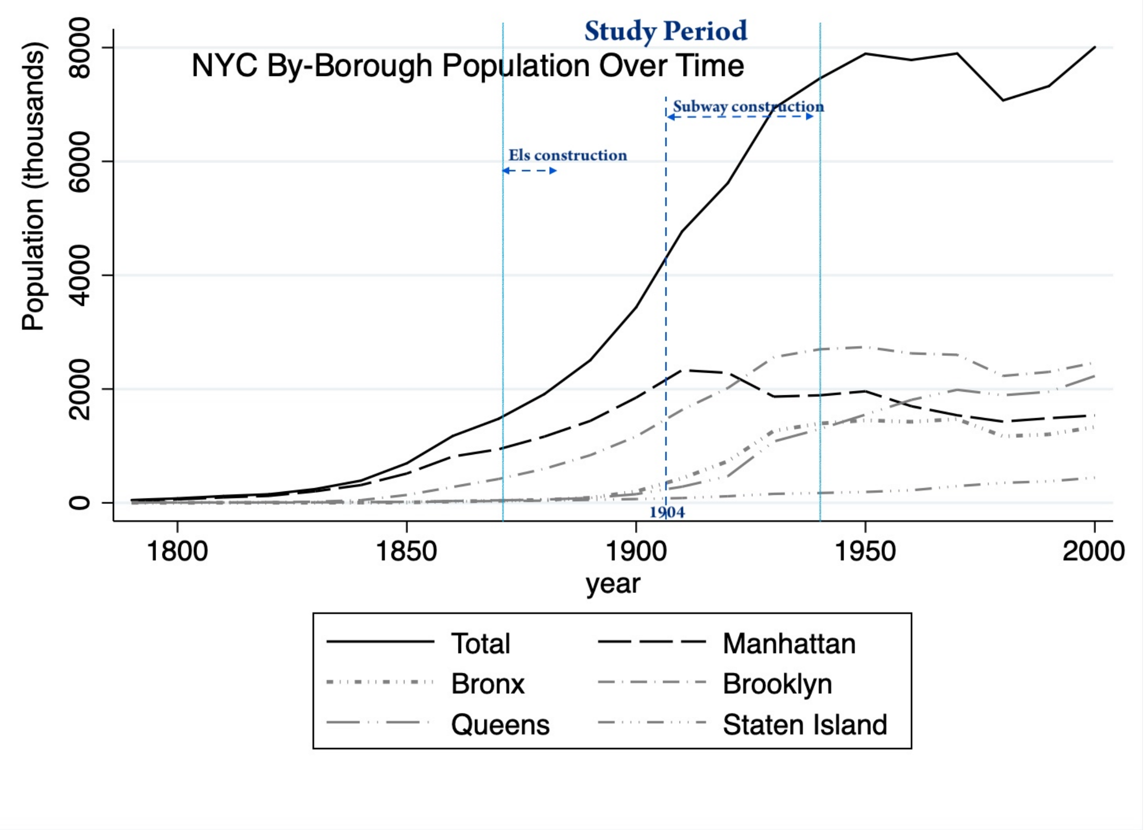
(b) Geographic Boundary of the Core and the Periphery of the City

1.3 New York City Background: 1870-1940

1.3.1 Population Growth

New York City was the largest city in the country at the beginning of the study period. Over the study period (1870-1940), the total population of the city increased from 1.48 million to 7.5 million.[7] During the study period, the total population in NYC (5 boroughs) experienced an astonishing growth with its peak population growth rate being 39% over a decade. However, beginning in the early twentieth century, Manhattan experienced the dramatic population loss when all outer boroughs were gaining population at an unprecedented rate (for example, between 1920 and 1930, Manhattan lost 18% of its population when the population in Queens and Bronx grew by 130% and 73% respectively).

Figure 12: Population Trend Over Time by Borough



[7] In 1898, through the consolidation of NYC, outer boroughs (Brooklyn, Bronx, Queens, and Staten Island) were incorporated into New York City. For my analysis, I always define the city as 5 boroughs throughout the study period.

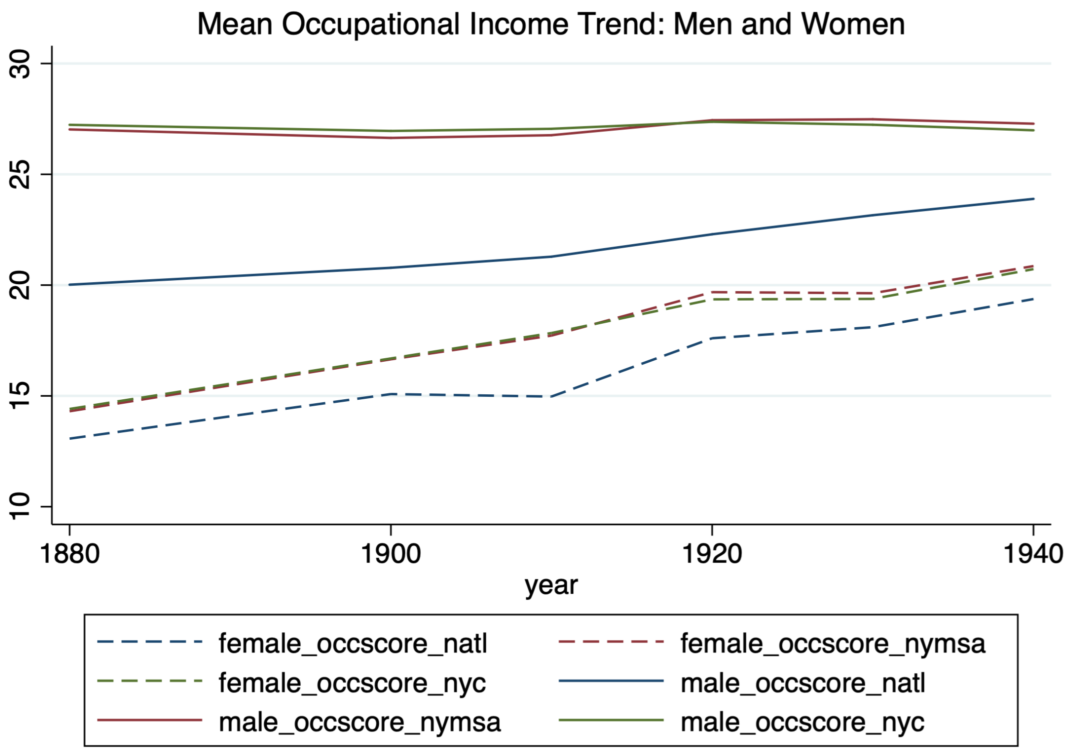
1.3.2 Income and Occupation Trends in NYC

NYC was growing in skill during the study period, as well as in population, and this growth in skill was occurring among almost all demographic groups. This aggregate skill growth matters for my analysis because it implies that *growth in skill in one neighborhood did not have to come at the expense of a reduction in skill in others*; the tide was rising and so no boat was forced to sink. However, skill growth in NYC was nowhere near as fast as population growth, and in some decades faltered slightly. New York was more skilled than the rest of the nation during the study period, but its advantage was eroding.

Figure 13 shows mean occupational income trend of all men and women aged between 16-60 with occupation over time at the varying geographic scope. Solid lines indicate men, whereas dotted lines indicate women; in terms of geography, the national average is in Blue, NYC's metro area average is in Red[8], and NYC average is in Green. Data reveal that men in NYC and NYC metro area had significantly higher mean occupational income than the rest of the country, but converged to the rest of the country over the 60 years. A similar pattern was observed for women but at a much smaller magnitude.

[8] A metro area is a region consisting of a large urban core together with surrounding communities that have a high degree of economic and social integration with the urban core. Since 1950, the Bureau of the Budget (later renamed the Office of Management and Budget, or OMB), has produced and continually updated standard delineations of metropolitan areas for the U.S. as a set of cities or towns. To delineate metro areas in pre-1950 samples (which is the case of all US census data that I use for the analysis), the general approach (used first by the creators of the 1940 PUMS and then by IPUMS for earlier samples) is to apply the 1950 OMB standards to historical statistics. This approach of applying the 1950 OMB standards to pre-1950 samples has merits as it reflects the evolution of population and economic integration between surrounding areas and the urban core over time.

Figure 13: Mean Occupational Income Trend: Men and Women



Source: US complete-count census records. All observations are aged between 16-60 with reported occupations.

1.4 Testing the Specific Propositions of Crabgrass Frontier

1.4.1 Did Population Grow in More Suburban Areas?

As the distance from the center increases (measured by the distance from the Battery which is the southern tip of Manhattan), the population density was declining during the study period. Table 11 and Figure 14 show that the population density gradient was negative and flattening. With the NTAs (neighborhoods in the city) as the units of observation, I regress log of population density as a function of the distance from the Battery to centroids of NTAs in the city. Regression results show that population gradient is negative and statistically significant, but starting from the peak of subway construction in the 1910s, the population density gradient was flattening significantly. The population density decreased as it gets further away from the center of the city. However, due to the transit infrastructure improvement, the population grew in more suburban areas and therefore the density gradient was flattening.[9]

The population grew in more suburban areas and Figure 17 shows this pattern over the study period. In 1880, only the center of the city and its adjacent areas were populated and areas further away from the center were largely unpopulated ("white shade" areas in Figure 17 indicates unpopulated areas, whereas the darker shades of the color red, the higher population density). However, starting from 1900, areas away from the center became populated and toward the end of the study period, in 1940, all neighborhoods in NYC became populated. Similar patterns are observed in Figure 16: in the beginning of study period (in 1880), only the areas close to the center were populated and areas further away from the center were largely unpopulated. Over time, areas relatively closer from the center became populated, and the slope of bivariate plots became flatter which imply that the population density declines less as the distance from the center increases. To the very end of study period (in 1940), basically all areas in the city became populated.[10]

The population density in places close to the center ("transit-hub") such as downtown and midtown Manhattan experienced dramatic losses, whereas "transit-spoke" neighborhoods such as upper Manhattan and Bronx, and Brooklyn were extensively gaining population.[11]

[9] This is consistent with the land use theory developed by Alonso (1964), Muth (1964), Mills (1967) which predicts that faster commuting times push up the demand for space in suburbs relative to central cities.

[10] In Figure 16, for $\ln(\textit{population density})$ (y-axis), I take the natural log of the total population in each NTA divided by the size of each NTA. With regards to the distance from center (x-axis), I measure the distance from the Battery, the southern tip of Manhattan in NYC, to centroids of each NTA in the city (measured in *kilometer*). I assign the value of $\ln(\textit{population density})$ to be nil for unpopulated NTAs with population of 0.

[11] In the Appendix, I map the Transit Access changes by decade drive by the elevated and subway construction by every decade during the study period. At the same geographic and time scale, I also map the new construction of residential-land use construction and commercial-land use construction by decade. Figures show that in places near the center ("transit-hub"), land became more dedicated for commercial use; whereas places far from the center but connected to the center ("transit-spoke"), land became more dedicated for residential use.

As in Figure 18, population density dramatically decreased in the center whereas the population increased substantially in surrounding areas of the city center. During the 1910s and 1920s, subway construction was at its peak through the Dual Contract period, and most neighborhoods in upper Manhattan, Brooklyn, and Bronx were experiencing a huge improvement in commuting transit access.[12]

Table 11: Population Density (with zeros)

	1870	1880	1900	1910	1920	1930	1940
Dist Battery	-0.0463***	-0.0984***	-0.0975***	-0.0898***	-0.0775***	-0.0467***	-0.0411***
	(0.00831)	(0.00831)	(0.00883)	(0.00852)	(0.00853)	(0.00614)	(0.00609)
Constant	3.167***	6.488***	8.260***	8.794***	8.736***	8.456***	8.348***
	(0.417)	(0.417)	(0.443)	(0.428)	(0.428)	(0.308)	(0.306)
N	195	195	195	195	195	195	195
R^2	0.139	0.421	0.387	0.365	0.300	0.230	0.191

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Dependent variable: $\ln(population\, density)$ normalized by the size of each NTA (measured in $kilometer^2$). Indepedent variable: distance from the Battery, the southern tip of Manhattan in NYC, to centroids of each NTA in the city (measured in *kilometer*). Here, when I take natural log of population density, I assign the value log(population density) to be nil for unpopulated NTAs with population of 0. 1870 (the first column) coefficient is less reliable as the availability of geographic information is extremely limited that identifying and harmonizing one's residential location at NTA level was more challenging than other years.

[12] Finally, in the 1920s, subfigure 18d shows that the huge population decline in upper east Manhattan and Harlem during this period. Harlem was predominantly occupied by Jewish and Italian in the 19th century. However, in the 1920s and 1930s, during the Great Migration, African-American residents arrived in large numbers and Harlem became the focus of the "Harlem Renaissance" and predominantly an African-American community.

Table 12: Population Density (without zeros)

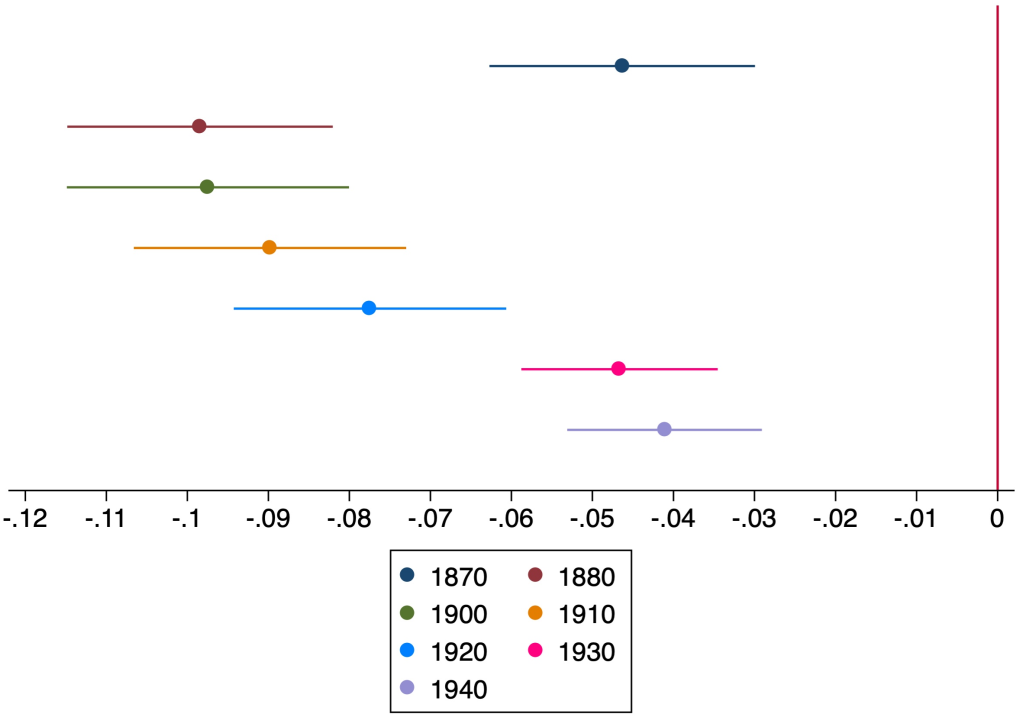
	1870	1880	1900	1910	1920	1930	1940
Dist Battery	-0.0691***	-0.0603***	-0.0630***	-0.0602***	-0.0569***	-0.0465***	-0.0361***
	(0.0151)	(0.00783)	(0.00631)	(0.00610)	(0.00633)	(0.00595)	(0.00542)
Constant	8.571***	8.107***	8.316***	8.523***	8.597***	8.480***	8.239***
	(0.581)	(0.264)	(0.281)	(0.284)	(0.303)	(0.299)	(0.270)
N	32	60	127	152	165	194	191
R^2	0.412	0.505	0.444	0.394	0.332	0.241	0.190

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Note: Dependent variable: $\ln(population\,density)$ normalized by the size of each NTA (measured in $kilometer^2$). Indepedent variable: distance from the Battery, the southern tip of Manhattan in NYC, to centroids of each NTA in the city (measured in $kilometer$). Here, when I take natural log of population density, I excluded unpopulated NTAs with population of 0. 1870 (the first column) coefficient is less reliable as the availability of geographic information is extremely limited that identifying and harmonizing one's residential location at NTA level was more challenging than other years.

Figure 14: Population Density Result Coefficients and Confidence Intervals (with zeros)



Note: When I take natural log of population density, I assign the value log(population density) to be nil for unpopulated NTAs with population of 0.

Figure 15: Population Density Result Coefficients and Confidence Intervals (without zeros)

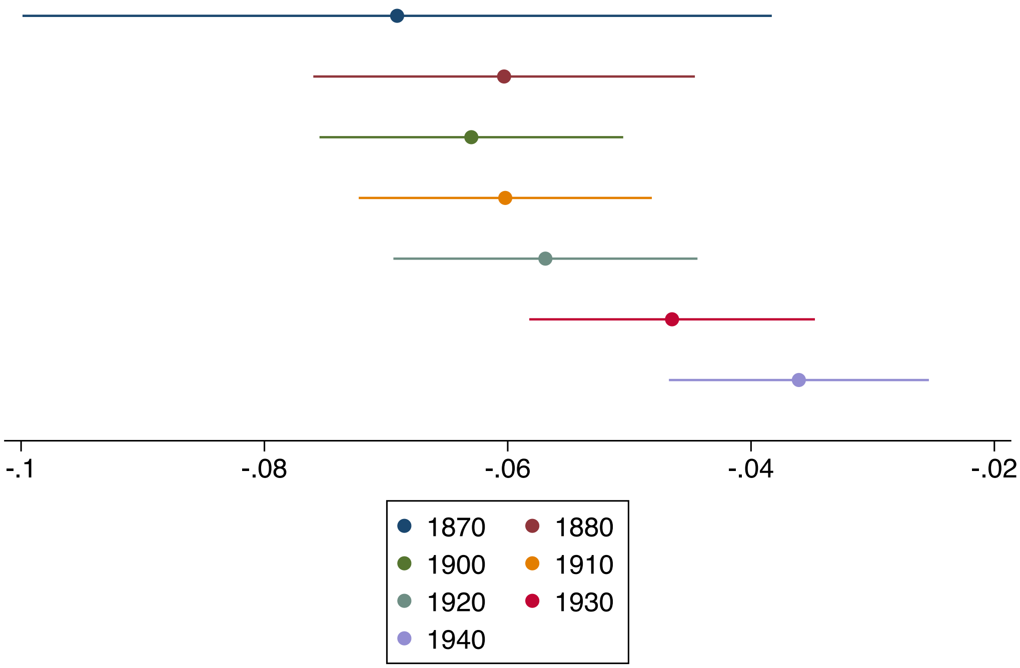
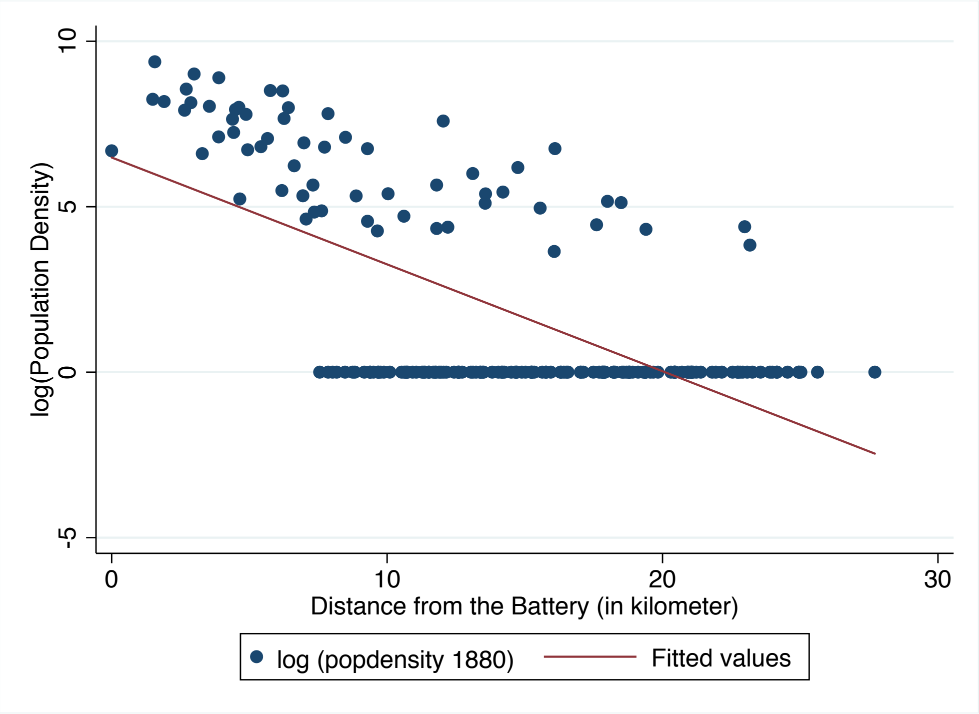


Figure 16: ln(*Population Density*) Against Distance From Center

(a) 1880



(b) 1900

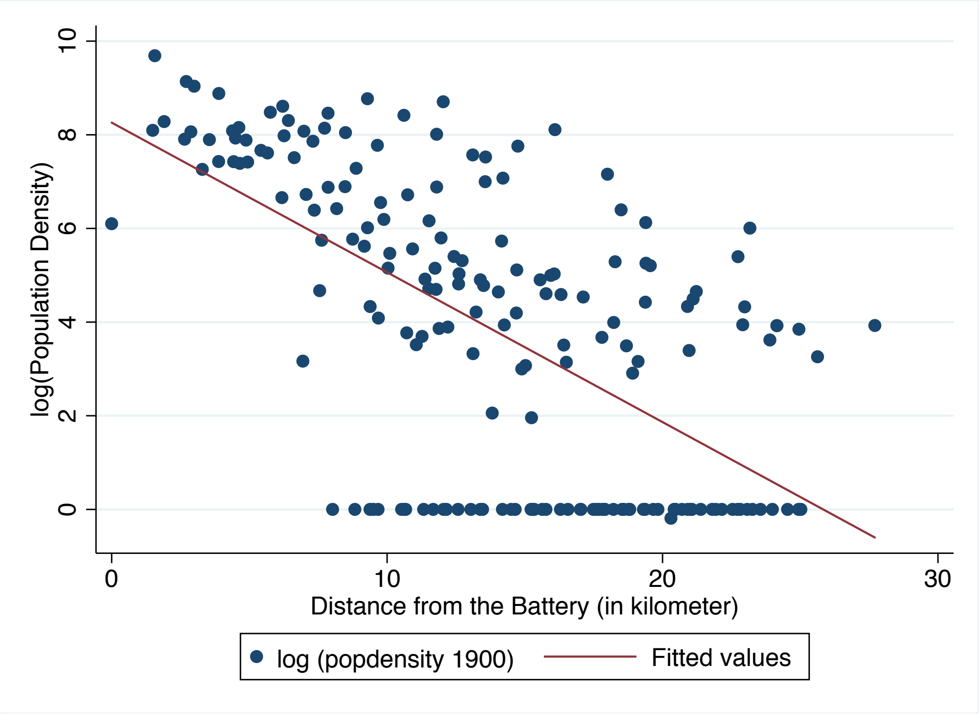
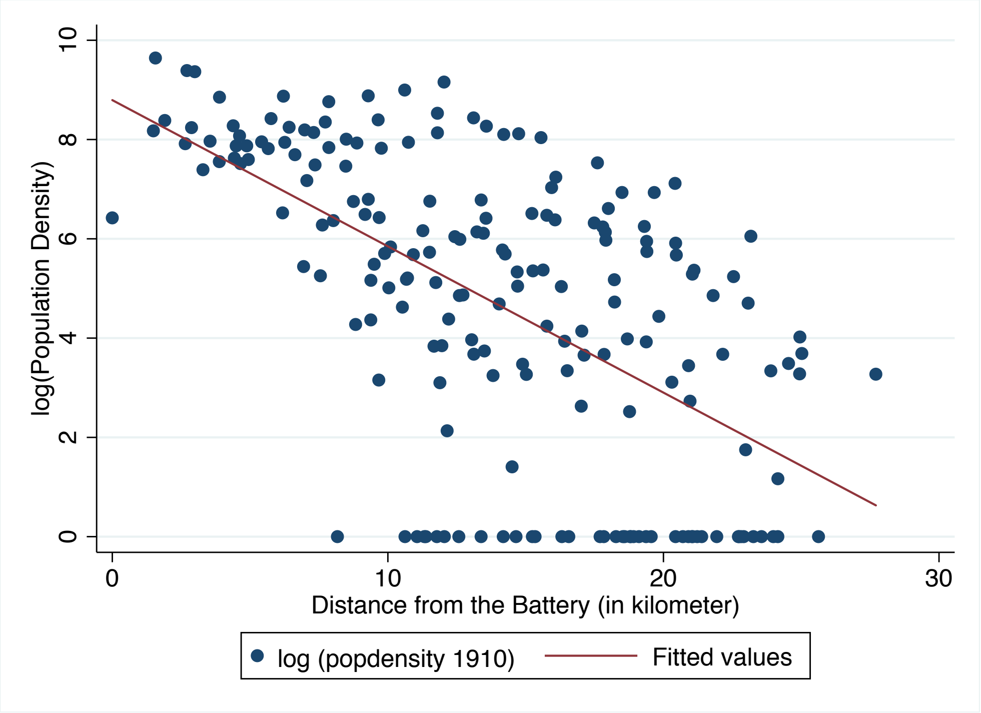


Figure 16: ln(*Population Density*) Against Distance From Center

(c) 1910



(d) 1920

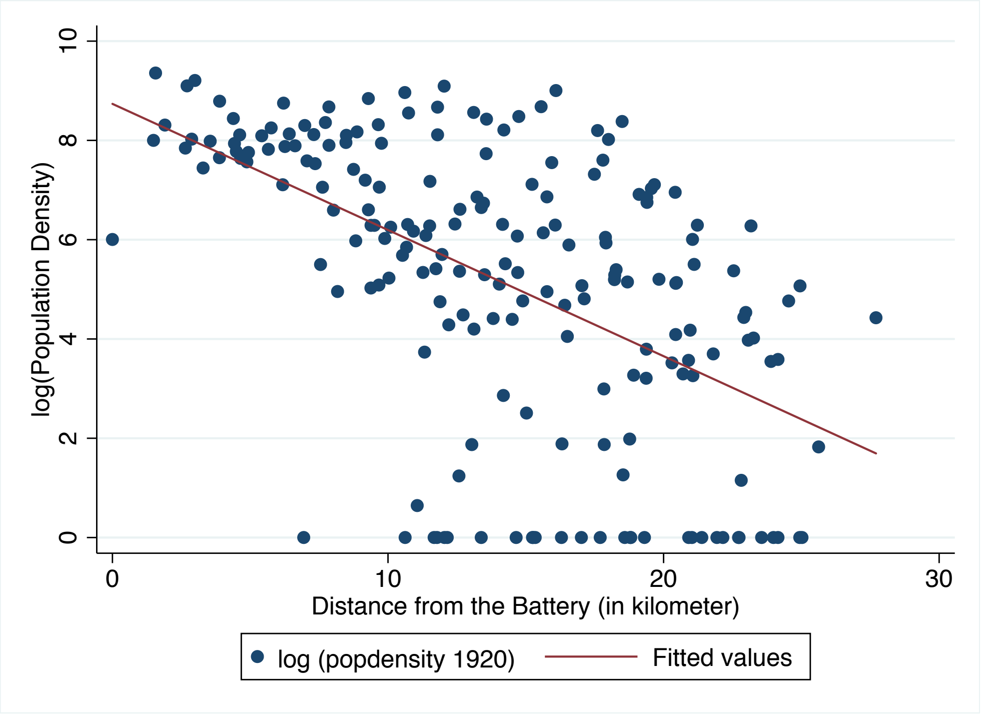
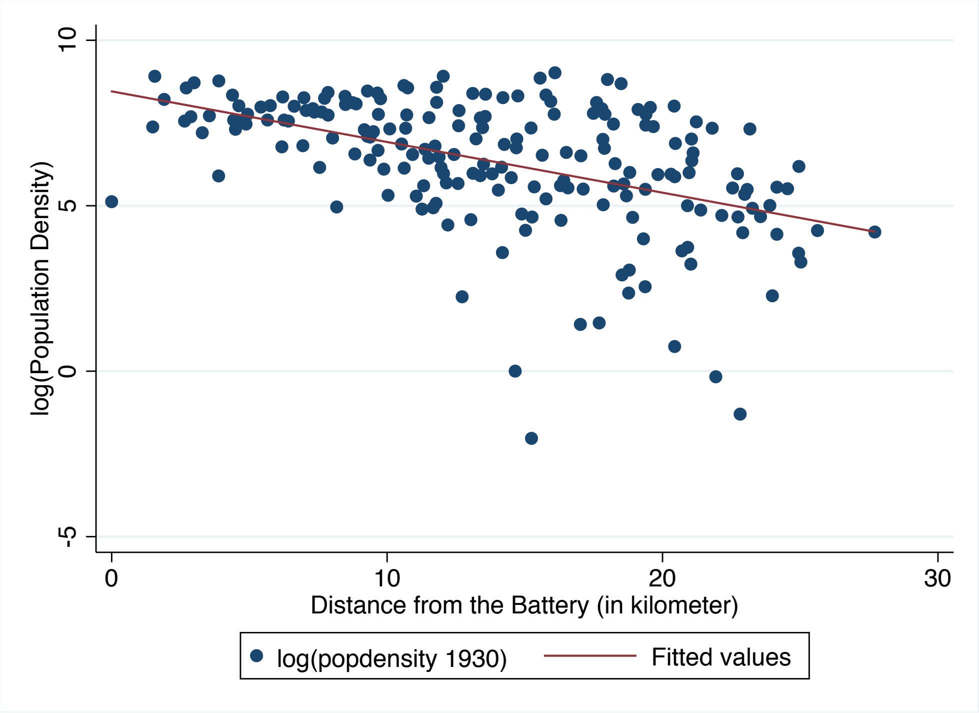


Figure 16: ln(*Population Density*) Against Distance From Center

(e) 1930



(f) 1940

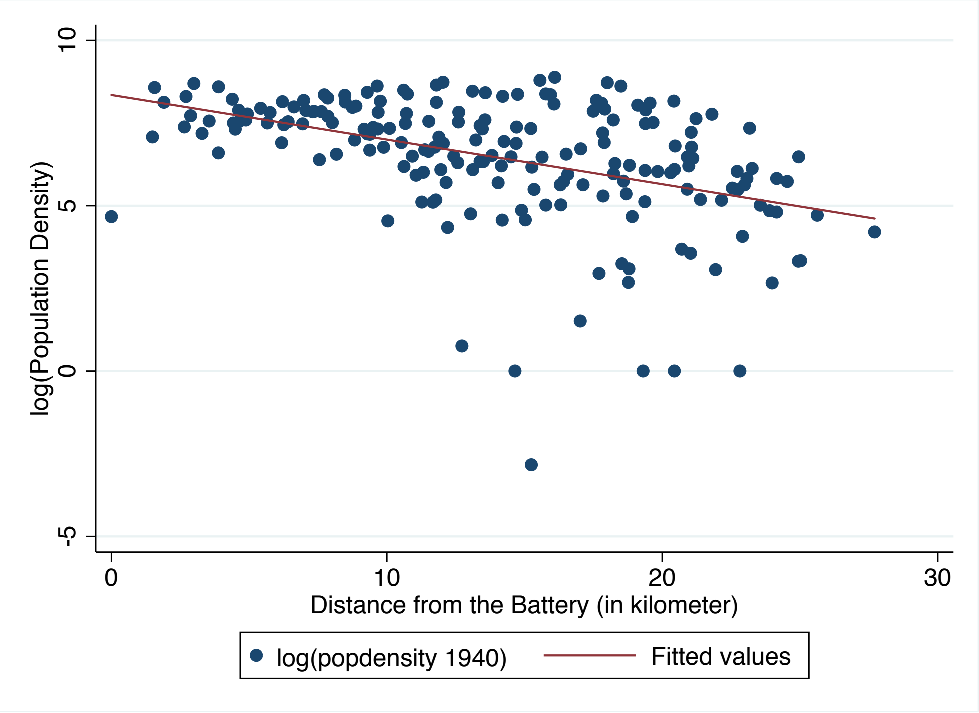
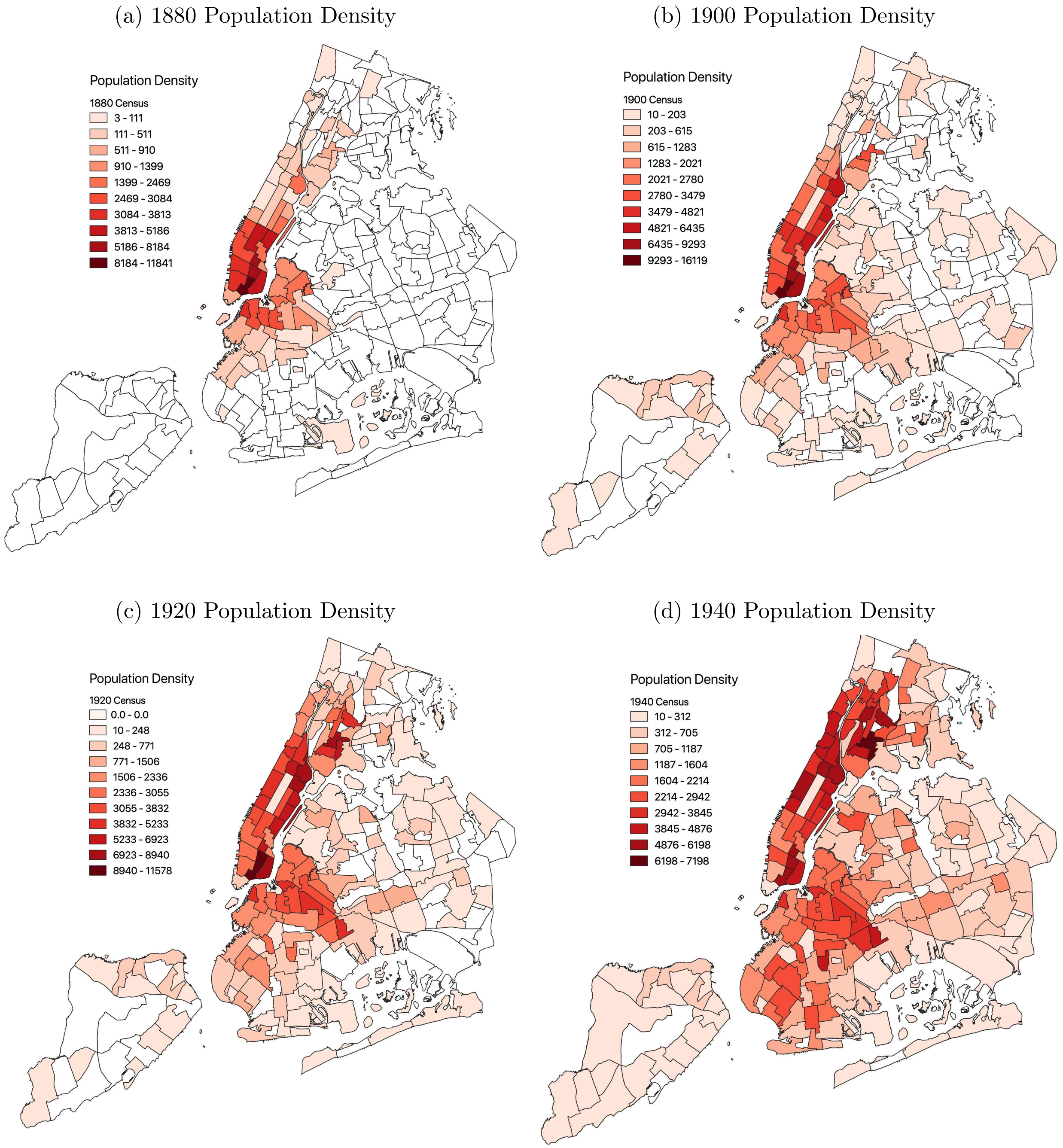


Figure 17: Population Density


(a) 1880 Population Density

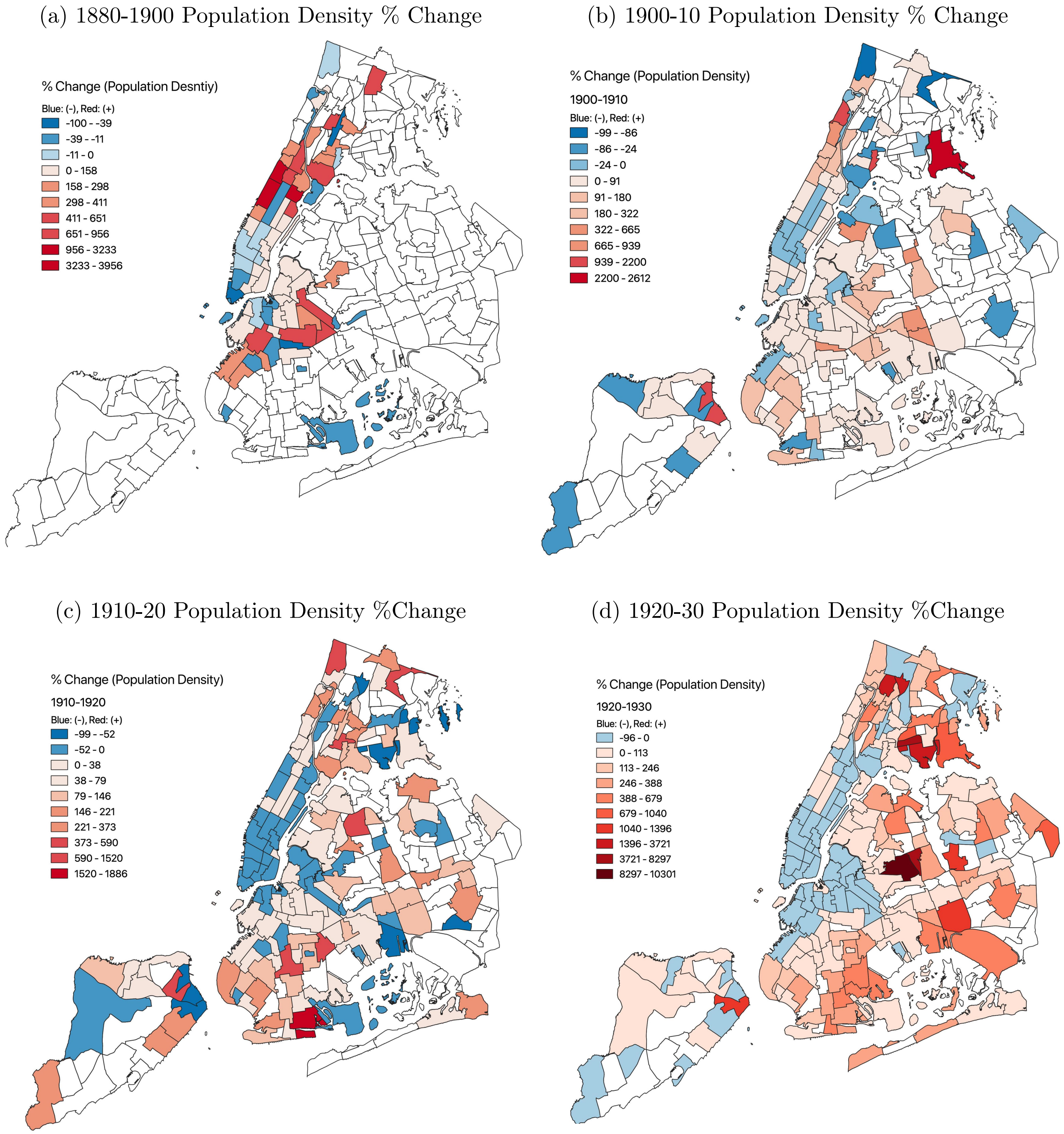
(b) 1900 Population Density

(c) 1920 Population Density

(d) 1940 Population Density

Note: The above figures show percent change of population density between two adjacent census periods. Source: Author's creation using the complete-count US Federal Demographic Census.

Figure 18: % Change of Population Density



(a) 1880-1900 Population Density % Change

(b) 1900-10 Population Density % Change

(c) 1910-20 Population Density %Change

(d) 1920-30 Population Density %Change

Note: The above figures show percent change of population density between two adjacent census periods. Source: Author's Creation using the complete-count US Federal Demographic Census.

1.4.2 Did the Rich Leave the Center of the City?

Jackson (1985) discusses the phenomenon in the 1850s NYC of the rich leaving the center of the city. He discusses the migration of the rich in the center of the city by quoting phrases concerning the 1850s New York such as "the desertion of the city by its men of wealth" and "many of the rich and prosperous are removing from the city, while the poor are pressing in."

If the popularly perceived pattern of the 1850s NYC had held true for my study period, the longitudinal database should reveal that leavers of the center of the city should be richer than the stayers. Therefore, among residents of the center of the city, I compare the occupational income of residents who later moved ("leavers") with that of those who stayed for the next decade ("stayers").

Throughout the period from 1880 to 1930, the longitudinal database shows that it was not the rich who left the center of the city. For example, Figures 19, 110, 111, 112 show that the mean occupational income of city-center leavers was lower than that of city-center stayers. In each NTA, blue shades (the darker blue, the poorer leavers) indicate leavers being poorer, whereas red shades (the darker red, the richer leavers) indicate leavers being richer than the stayers. At the center of the city, throughout the years between 1880 and 1930, in Figures 19, 110, 111, 112, the core of the city being consistently blue indicates that it was not the rich who left the center of the city—in fact, the leavers had lower mean occupational income than the center-stayers.

Regression results also show that it was not the rich who left the center of the city. I run logistic regression (also called as a logit model) to model the log odds of individuals' leaving the city relative to staying in the city in the later period, using the longitudinal data of individuals during the study period. The outcome of interest is identifying factors that explain whether individuals living in the core of the city in the early period leaves or stays the city boundary (5 boroughs) in the subsequent period. The predictor variables of interest are occupational income, nativity, race, and age.

Regression results in Tables 13, 15, 17, 19, 111 show that as occupational income increases, people who lived in the city center in the earlier period were less likely to leave the city. In terms of the nativity, being foreign-born relative to native-born with both native parents decreases the log odds of leaving the city center — this may be partially due to ethnic enclaves in Lower East Side of Manhattan near the city center. In terms of race, being non-white relative to white increases the log odds of leaving the city center and the degree of relative log odds across race differ; however, considering that the majority of residents in New York were white, this may be interpreted with caution. Finally, older people are more likely to leave throughout the study period.

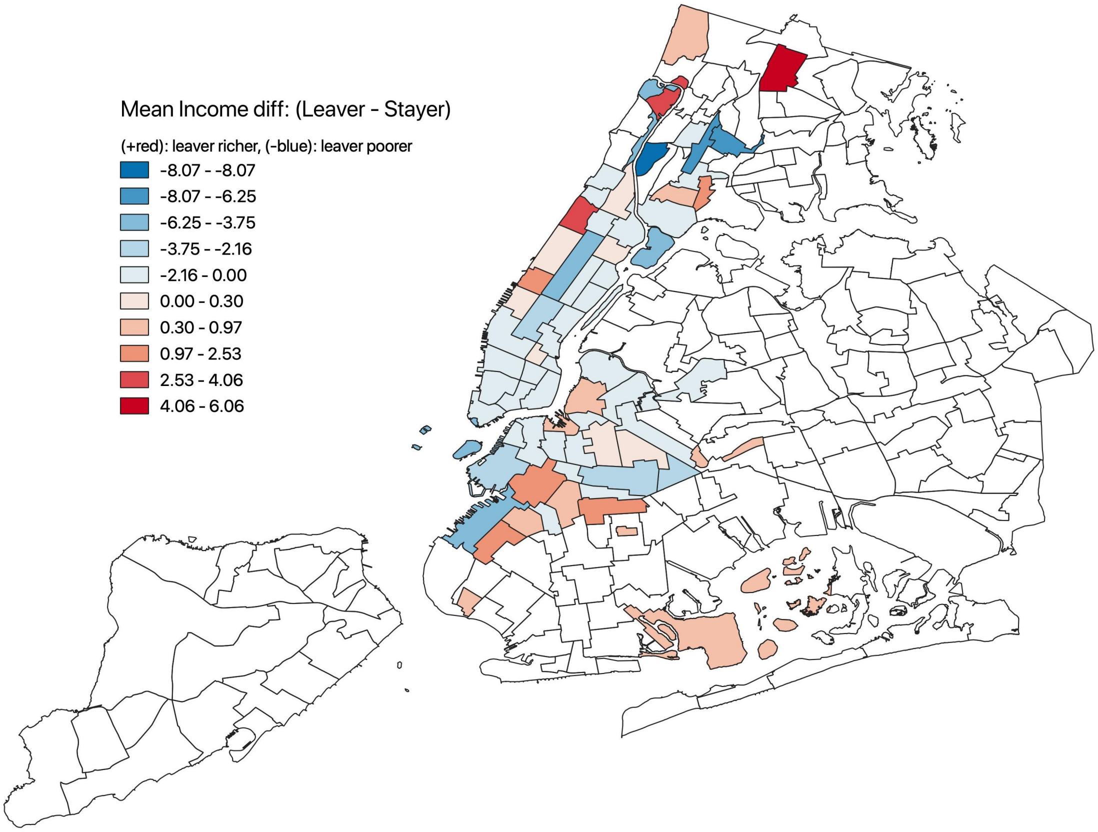
While the regression results in Tables 13, 15, 17, 19, 111 look at extensive margin of leaving or staying in the city among people who lived at the core of the city in the earlier period, Tables 14, 16, 18, 110, 112, 121 look at whether flows to the metro area may have been different from flows to outside the metro area (e.g. the city core leavers migrating to places like California that are strictly outside NYC metro area but in the country boundary), and flows within the city (i.e. the city core leavers migrating to the periphery of the city) relative the people who stayed in the city core both periods.

Regression results in Tables 13, 15, 17, 19, 111 show that relative to people people who stayed in the city center, as occupational income increases, people who lived in the city center in the earlier period were less likely to leave the city at every migration scale — leaving to NYC metro area, leaving to outside NYC metro area, moving to the periphery of the city and this holds up until the Great Depression. In terms of the nativity, being a foreign-born relative to native-born with both native parents increases the log odds of leaving the city center. Finally, older people were less likely to leave the city center.

The people whom public opinion perceive to be richer may be older and more likely to be native whites of native parentage (and therefore they are more “prestigious”). So, the public perception may have been that the leavers had higher social status, not that they had a higher income. However, the occupational income measure that I use for the analysis only depends on one’s occupation, and it does not reflect factors such as one’s race, nativity, or age, and such factors may have played more roles in determining one’s income. This makes the accuracy of income measures more crucial. If there was wage discrimination against who people who were not old or not native whites with native parentage, then leavers of the core may have been richer. Not necessarily more skilled (in terms of occupation), but richer. Considering that income tends to increase with age, to the contemporary observers, the relatively old people's leaving the city center may have been interpreted as *"the desertion of the city by its men of wealth."*

Relatedly, Section 1.4.4.2 discusses decomposition of various flows of the core of the city including the relative income difference between leavers and stayers as well as the corresponding relative magnitudes of those flows at the neighborhood level. The people from the core who left the metropolitan area were richer than the people from the periphery who left the metropolitan area, and poorer people from outside NYC metro area migrated to the core over time, making the relative income at the core to decrease.

Figure 19: Neighborhood-level Mean Income Differences between Leavers and Stayers, 1880-1900



Note: Blue shades mean leavers' mean occupational income was lower than stayers, whereas red shades mean leavers' mean occupational income was higher than stayers in the Year 1880.

Table 13: Logit Results between Leavers and Stayers at the City level: 1880-1900 Males

[City Leavers]				
Occupational income	-0.00517**	-0.0101***	-0.0106***	-0.0121***
	(0.00199)	(0.00208)	(0.00212)	(0.00215)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-1.350***	-1.357***	-1.295***
		(0.205)	(0.205)	(0.206)
Native born: mother foreign, father native		-1.059***	-1.072***	-1.006***
		(0.284)	(0.285)	(0.285)
Native born: both parents foreign		-1.488***	-1.501***	-1.422***
		(0.0661)	(0.0670)	(0.0696)
Foreign-born		-0.177***	-0.191***	-0.206***
		(0.0526)	(0.0537)	(0.0539)
Race				
White			-	-
Black			-0.229	-0.226
			(0.180)	(0.180)
Chinese			0.156	0.218
			(0.377)	(0.378)
Age				0.00979***
				(0.00241)
[City Stayers]	(base outcome)			
N	9920	9920	9920	9920

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 14: Multinomial Logit Results between Leavers and Stayers at Neighborhood Level: 1880-1900 males

[City Core Stayers]: baseline comparison group				
[City leavers & NYC metro area stayers]				
Occupational income	-0.0152**	-0.0216***	-0.0223***	-0.0171**
	(0.00551)	(0.00561)	(0.00578)	(0.00592)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-1.292**	-1.295**	-1.512***
		(0.444)	(0.444)	(0.448)
Native born: mother foreign, father native		-0.472	-0.469	-0.709
		(0.717)	(0.717)	(0.719)
Native born: both parents foreign		-1.220***	-1.218***	-1.515***
		(0.169)	(0.171)	(0.180)
Foreign-born		-0.771***	-0.779***	-0.732***
		(0.151)	(0.152)	(0.152)
Race				
White			-	-
Black			0.789	0.762
			(1.064)	(1.064)
Chinese			13.81	13.56
			(772.0)	(767.9)
Age				-0.0338***
				(0.00635)
[City & NYC metro area Leavers]				
Occupational income	-0.00909*	-0.0151***	-0.0137**	-0.0105*
	(0.00430)	(0.00444)	(0.00453)	(0.00468)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-2.034***	-2.016***	-2.157***
		(0.348)	(0.348)	(0.351)
Native born: mother foreign, father native		-1.214	-1.174	-1.340*
		(0.624)	(0.624)	(0.626)
Native born: both parents foreign		-1.813***	-1.772***	-1.976***
		(0.141)	(0.142)	(0.151)
Foreign-born		-0.282*	-0.251	-0.220
		(0.128)	(0.129)	(0.129)
Race				
White			-	-
Black			1.598	1.573
			(1.012)	(1.012)
Chinese			13.69	13.52
			(772.0)	(767.9)
Age				-0.0224***
				(0.00513)

Continued

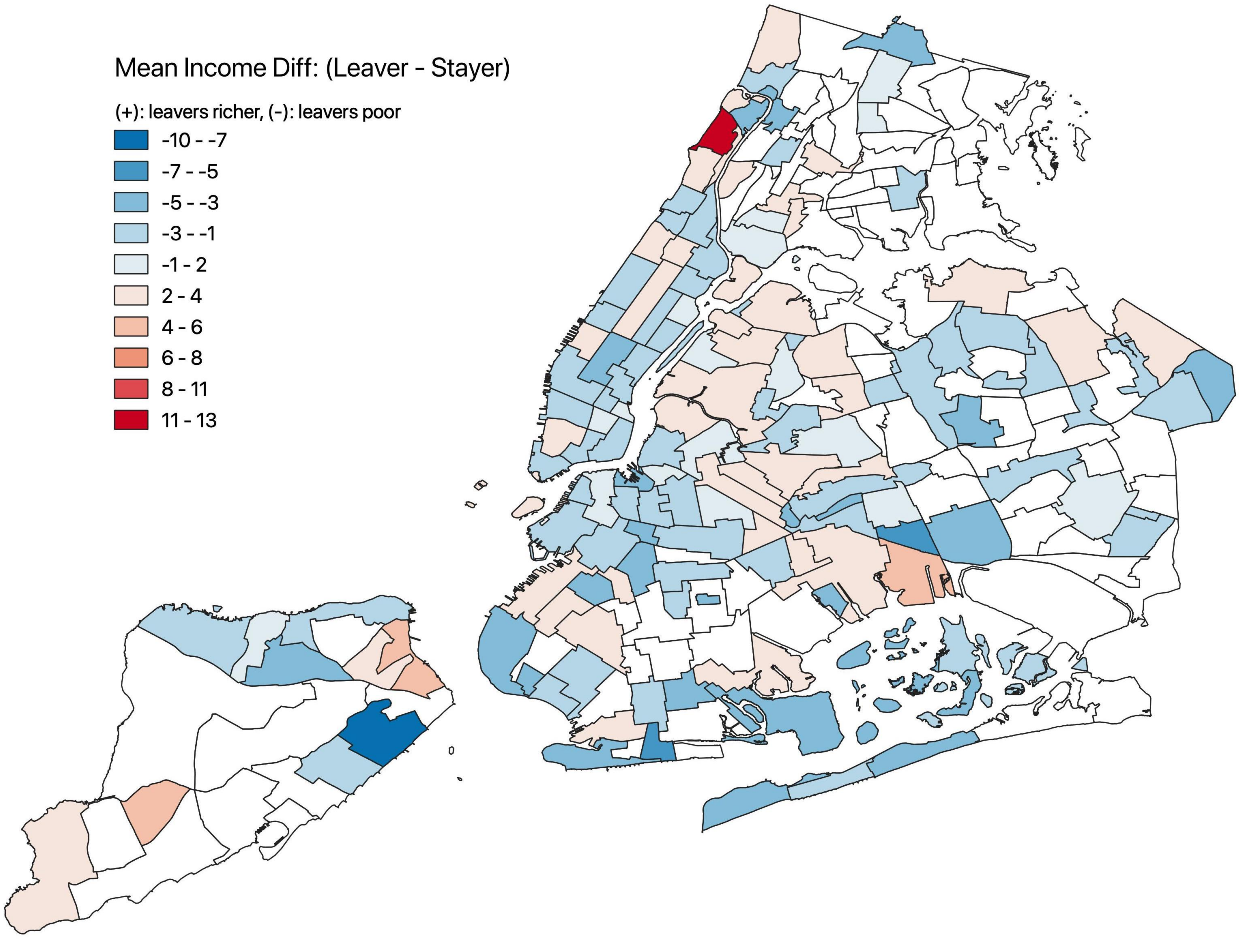
Table 14: Multinomial Logit Results between Leavers and Stayers at Neighborhood Level: 1880-1900 males (cont.)

[City Core Stayers]: baseline comparison group				
[City Stayer & City Core Leavers]				
Occupational income	-0.00533	-0.00671	-0.00475	0.00112
	(0.00451)	(0.00461)	(0.00471)	(0.00486)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-0.685*	-0.659	-0.911**
		(0.345)	(0.345)	(0.350)
Native born: mother foreign, father native		-0.0412	0.0139	-0.264
		(0.628)	(0.628)	(0.632)
Native born: both parents foreign		-0.274	-0.217	-0.557***
		(0.143)	(0.144)	(0.154)
Foreign-born		-0.174	-0.128	-0.0714
		(0.134)	(0.135)	(0.136)
Race				
White			-	-
Black			1.865	1.836
			(1.019)	(1.020)
Chinese			13.70	13.42
			(772.0)	(767.9)
Age				-0.0396***
				(0.00543)
N	9920	9920	9920	9920

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 110: Neighborhood-level Mean Income Differences between Leavers and Stayers, 1900-1910


Mean Income Diff: (Leaver - Stayer)
(+): leavers richer, (-): leavers poor
-10 - -7
-7 - -5
-5 - -3
-3 - -1
-1 - 2
2 - 4
4 - 6
6 - 8
8 - 11
11 - 13

Note: Blue shades mean leavers' mean occupational income was lower than stayers, whereas red shades mean leavers' mean occupational income was higher than stayers in the Year 1900.

Table 15: Logit Results between Leavers and Stayers at the City level: 1900-1910 males

[City Leavers]				
Occupational income	-0.0105***	-0.0103***	-0.00978***	-0.0112***
	(0.00154)	(0.00160)	(0.00162)	(0.00164)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-1.377***	-1.373***	-1.318***
		(0.145)	(0.146)	(0.146)
Native born: mother foreign, father native		-1.222***	-1.218***	-1.173***
		(0.221)	(0.221)	(0.221)
Native born: both parents foreign		-1.003***	-1.000***	-0.981***
		(0.0532)	(0.0551)	(0.0552)
Foreign-born		-0.343***	-0.350***	-0.371***
		(0.0457)	(0.0481)	(0.0483)
Race				
White			-	-
Black			0.0324	0.0228
			(0.129)	(0.129)
Chinese			2.010***	1.983***
			(0.365)	(0.366)
Japanese			12.54	12.53
			(526.8)	(527.8)
Age				0.0106***
				(0.00148)
[City Stayers]	(base outcome)			
N	19085	19085	19085	19085

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 16: Multinomial Logit Results between Leavers and Stayers at Neighborhood Level: 1900-1910 males

[City Core Stayers]: baseline comparison group				
[City leavers & NYC metro area stayers]				
Occupational income	-0.0193***	-0.0197***	-0.0202***	-0.0171***
	(0.00362)	(0.00370)	(0.00376)	(0.00379)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-0.802*	-0.825*	-0.956**
		(0.321)	(0.322)	(0.323)
Native born: mother foreign, father native		-0.133	-0.159	-0.253
		(0.458)	(0.459)	(0.461)
Native born: both parents foreign		-0.209	-0.235	-0.284*
		(0.128)	(0.132)	(0.133)
Foreign-born		-0.140	-0.171	-0.123
		(0.109)	(0.114)	(0.114)
Race				
White			-	-
Black			-0.252	-0.236
			(0.314)	(0.314)
Chinese			13.60	13.90
			(483.6)	(548.2)
Japanese			-0.0292	-0.0355
			(6027.3)	(6776.7)
Age				-0.0244***
				(0.00369)
[City & NYC metro area Leavers]				
Occupational income	-0.0262***	-0.0252***	-0.0249***	-0.0238***
	(0.00256)	(0.00263)	(0.00267)	(0.00269)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-1.455***	-1.459***	-1.511***
		(0.226)	(0.227)	(0.227)
Native born: mother foreign, father native		-1.230**	-1.234**	-1.265***
		(0.375)	(0.376)	(0.377)
Native born: both parents foreign		-0.627***	-0.632***	-0.655***
		(0.0981)	(0.101)	(0.102)
Foreign-born		-0.114	-0.132	-0.116
		(0.0834)	(0.0874)	(0.0876)
Race				
White			-	-
Black			-0.0562	-0.0552
			(0.237)	(0.237)
Chinese			14.80	15.07
			(483.6)	(548.2)
Japanese			15.22	15.45
			(4208.0)	(4690.4)
Age				-0.00908***
				(0.00271)

Continued

Table 16: Multinomial Logit Results between Leavers and Stayers at Neighborhood Level: 1900-1910 males (cont.)

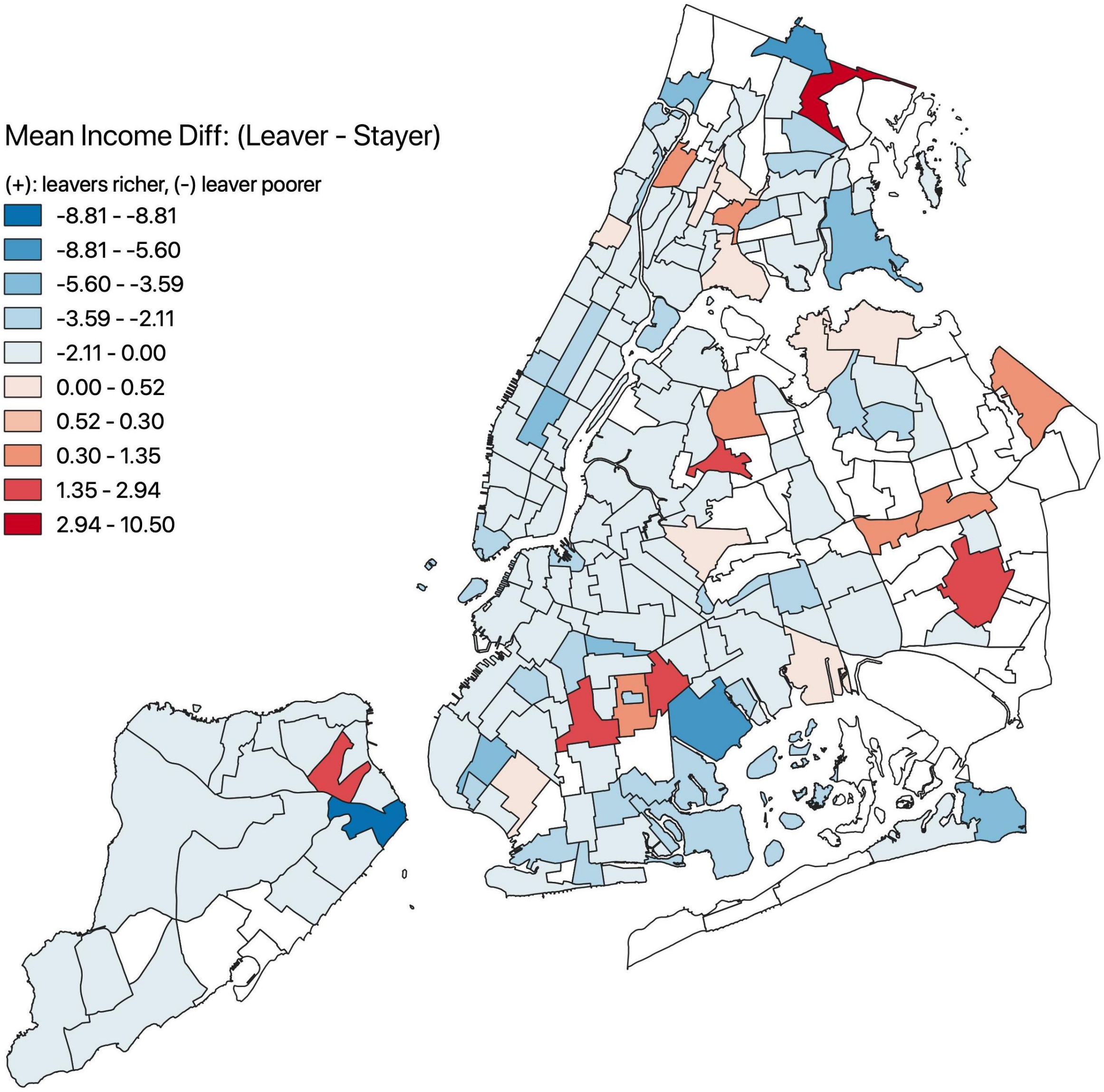
`[City Core Stayers]:  baseline comparison group`

[City Stayer & City Core Leavers]				
Occupational income	-0.0193***	-0.0188***	-0.0191***	-0.0157***
	(0.00271)	(0.00281)	(0.00285)	(0.00288)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		0.0217	0.0101	-0.134
		(0.221)	(0.222)	(0.223)
Native born: mother foreign, father native		0.239	0.226	0.122
		(0.366)	(0.367)	(0.370)
Native born: both parents foreign		0.541***	0.528***	0.475***
		(0.104)	(0.108)	(0.108)
Foreign-born		0.295**	0.279**	0.332***
		(0.0911)	(0.0954)	(0.0958)
Race				
White			-	-
Black			-0.133	-0.114
			(0.260)	(0.261)
Chinese			12.91	13.22
			(483.6)	(548.2)
Japanese			-0.0173	-0.0185
			(4729.7)	(5291.0)
Age				-0.0271***
				(0.00290)
N	19100	19100	19100	19100

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 111: Neighborhood-level Mean Income Differences between Leavers and Stayers, 1910-1920



Note: Blue shades mean leavers' mean occupational income was lower than stayers, whereas red shades mean leavers' mean occupational income was higher than stayers in the Year 1910.

Table 17: Logit Results between Leavers and Stayers at the City level: 1910-1920 males

[City Leavers]				
Occupational income	-0.0119***	-0.0116***	-0.0111***	-0.0127***
	(0.000788)	(0.000814)	(0.000824)	(0.000835)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-1.522***	-1.498***	-1.465***
		(0.0807)	(0.0810)	(0.0811)
Native born: mother foreign, father native		-1.689***	-1.665***	-1.628***
		(0.122)	(0.122)	(0.122)
Native born: both parents foreign		-1.192***	-1.170***	-1.159***
		(0.0309)	(0.0318)	(0.0318)
Foreign-born		-0.214***	-0.199***	-0.204***
		(0.0235)	(0.0247)	(0.0247)
Race				
White			-	-
Black			0.215**	0.202**
			(0.0703)	(0.0703)
Chinese			1.664***	1.585***
			(0.150)	(0.150)
Japanese			2.114*	2.131*
			(1.035)	(1.036)
Age				0.00942***
				(0.000756)
N	65336	65336	65336	65336

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 18: Multinomial Logit Results between Leavers and Stayers at Neighborhood Level: 1910-1920 males

[City Core Stayers]: baseline comparison group				
[City leavers & NYC metro area stayers]				
Occupational income	-0.0169***	-0.0161***	-0.0171***	-0.0132***
	(0.00182)	(0.00186)	(0.00189)	(0.00191)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-0.821***	-0.873***	-0.954***
		(0.152)	(0.152)	(0.153)
Native born: mother foreign, father native		-1.008***	-1.060***	-1.147***
		(0.230)	(0.231)	(0.231)
Native born: both parents foreign		-0.509***	-0.564***	-0.585***
		(0.0668)	(0.0682)	(0.0683)
Foreign-born		0.0225	-0.0318	-0.0188
		(0.0550)	(0.0568)	(0.0569)
Race				
White			-	-
Black			-0.747***	-0.722***
			(0.206)	(0.206)
Chinese			-0.854*	-0.686
			(0.416)	(0.417)
Japanese			-0.343	-0.403
			(1406.6)	(1404.9)
Age				-0.0221***
				(0.00179)
[City & NYC metro area Leavers]				
Occupational income	-0.0240***	-0.0216***	-0.0207***	-0.0182***
	(0.00126)	(0.00131)	(0.00132)	(0.00134)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-1.868***	-1.827***	-1.878***
		(0.114)	(0.115)	(0.115)
Native born: mother foreign, father native		-2.007***	-1.967***	-2.020***
		(0.169)	(0.169)	(0.170)
Native born: both parents foreign		-1.282***	-1.242***	-1.255***
		(0.0494)	(0.0508)	(0.0509)
Foreign-born		0.116**	0.150***	0.157***
		(0.0406)	(0.0423)	(0.0424)
Race				
White			-	-
Black			0.328*	0.342**
			(0.127)	(0.128)
Chinese			1.008***	1.107***
			(0.226)	(0.226)
Japanese			14.14	14.09
			(933.6)	(930.9)
Age				-0.0138***
				(0.00128)

Continued

Table 18: Multinomial Logit Results between Leavers and Stayers at Neighborhood Level: 1910-1920 males (cont.)

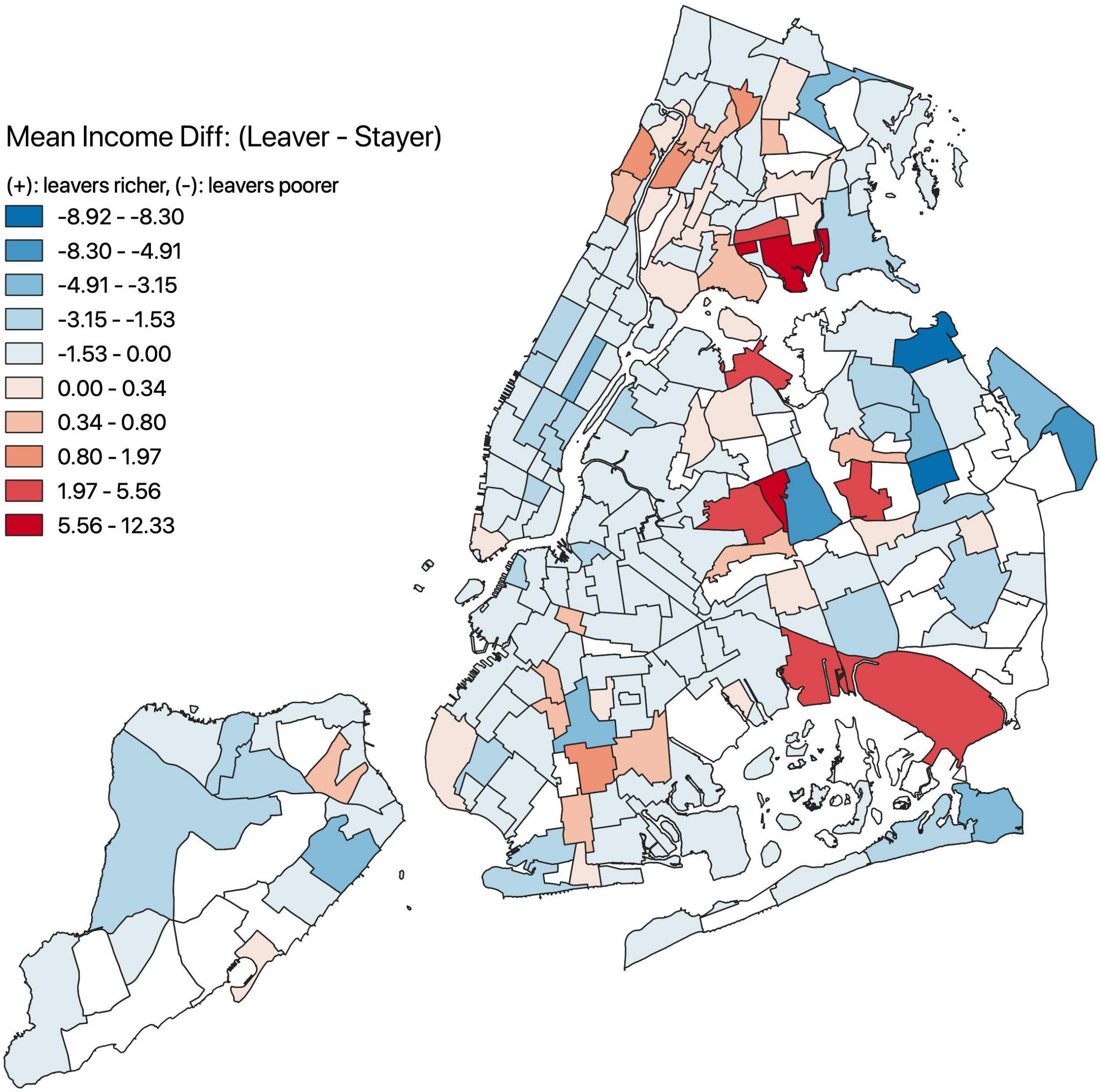
`[City Core Stayers]:  baseline comparison group`

[City Stayer & City Core Leavers]				
Occupational income	-0.0144***	-0.0121***	-0.0120***	-0.00648***
	(0.00129)	(0.00133)	(0.00134)	(0.00137)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-0.189	-0.182	-0.294**
		(0.0981)	(0.0987)	(0.0996)
Native born: mother foreign, father native		-0.168	-0.163	-0.284*
		(0.138)	(0.138)	(0.139)
Native born: both parents foreign		0.0641	0.0721	0.0395
		(0.0496)	(0.0511)	(0.0514)
Foreign-born		0.412***	0.422***	0.441***
		(0.0434)	(0.0451)	(0.0453)
Race				
White			-	-
Black			0.0350	0.0748
			(0.138)	(0.138)
Chinese			-1.088***	-0.842**
			(0.285)	(0.286)
Japanese			12.08	12.01
			(933.6)	(930.9)
Age				-0.0313***
				(0.00134)
N	65336	65336	65336	65336

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 112: Neighborhood-level Mean Income Differences between Leavers and Stayers, 1920-1930



Note: Blue shades mean leavers' mean occupational income was lower than stayers, whereas red shades mean leavers' mean occupational income was higher than stayers in the Year 1920.

Table 19: Logit Results between Leavers and Stayers at the City level: 1920-1930 males

[City Leavers]				
Occupational income	-0.0120***	-0.00744***	-0.00675***	-0.00976***
	(0.000510)	(0.000535)	(0.000540)	(0.000548)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-1.340***	-1.327***	-1.316***
		(0.0396)	(0.0397)	(0.0398)
Native born: mother foreign, father native		-1.302***	-1.288***	-1.249***
		(0.0574)	(0.0575)	(0.0577)
Native born: both parents foreign		-1.332***	-1.317***	-1.301***
		(0.0170)	(0.0173)	(0.0173)
Foreign-born		0.310***	0.319***	0.245***
		(0.0132)	(0.0135)	(0.0137)
Race				
White			-	-
Black			0.196***	0.178***
			(0.0361)	(0.0361)
Chinese			1.210***	1.180***
			(0.124)	(0.124)
Japanese			0.887**	0.896**
			(0.287)	(0.287)
Age				0.0175***
				(0.000505)
[City Stayers]	(base outcome)			
N	152589	152589	152589	152589

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 110: Multinomial Logit Results between Leavers and Stayers at Neighborhood Level: 1920-1930 males

[City Core Stayers]: baseline comparison group				
[City leavers & NYC metro area stayers]				
Occupational income	-0.0129***	-0.00998***	-0.0116***	-0.00765***
	(0.00108)	(0.00110)	(0.00112)	(0.00114)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-0.765***	-0.827***	-0.842***
		(0.0676)	(0.0678)	(0.0679)
Native born: mother foreign, father native		-0.685***	-0.750***	-0.794***
		(0.0972)	(0.0974)	(0.0976)
Native born: both parents foreign		-0.672***	-0.740***	-0.752***
		(0.0346)	(0.0350)	(0.0351)
Foreign-born		0.331***	0.273***	0.359***
		(0.0308)	(0.0312)	(0.0317)
Race				
White			-	-
Black			-1.175***	-1.152***
			(0.106)	(0.106)
Chinese			-0.874**	-0.843**
			(0.271)	(0.271)
Japanese			-1.531	-1.552
			(1.118)	(1.118)
Age				-0.0206***
				(0.00110)
[City & NYC metro area Leavers]				
Occupational income	-0.0244***	-0.0182***	-0.0179***	-0.0163***
	(0.000785)	(0.000817)	(0.000823)	(0.000837)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-1.842***	-1.845***	-1.848***
		(0.0570)	(0.0572)	(0.0572)
Native born: mother foreign, father native		-1.785***	-1.788***	-1.796***
		(0.0832)	(0.0833)	(0.0834)
Native born: both parents foreign		-1.450***	-1.453***	-1.452***
		(0.0263)	(0.0268)	(0.0269)
Foreign-born		0.548***	0.544***	0.566***
		(0.0231)	(0.0235)	(0.0238)
Race				
White			-	-
Black			-0.0784	-0.0626
			(0.0566)	(0.0567)
Chinese			0.124	0.140
			(0.164)	(0.164)
Japanese			0.569	0.565
			(0.521)	(0.520)
Age				-0.00523***
				(0.000799)

Continued

Table 110: Multinomial Logit Results between Leavers and Stayers at Neighborhood Level: 1920-1930 males (cont.)

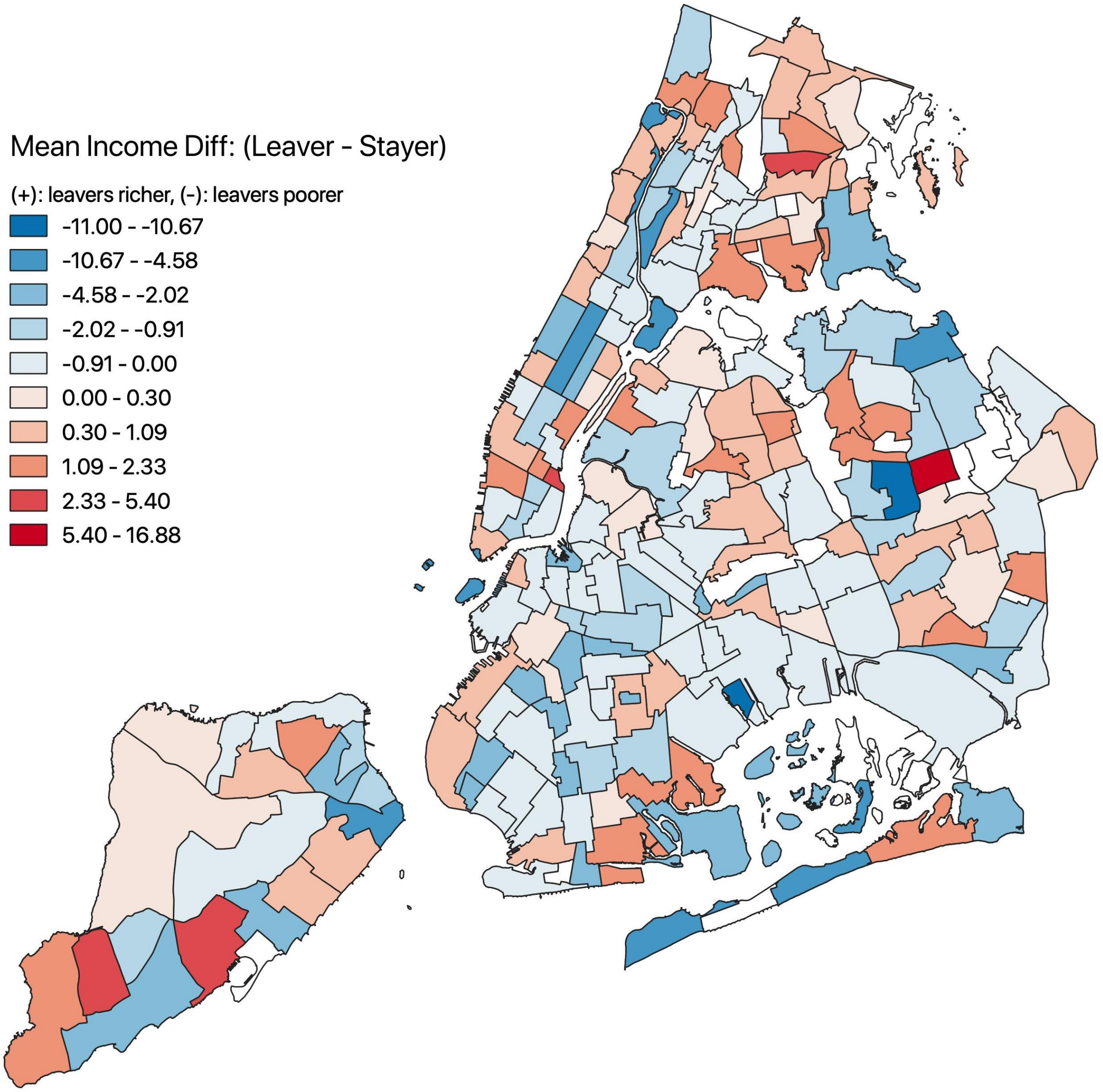
`[City Core Stayers]:  baseline comparison group`

[City Stayer & City Core Leavers]				
Occupational income	-0.0131***	-0.0116***	-0.0126***	-0.00629***
	(0.000730)	(0.000754)	(0.000759)	(0.000779)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-0.269***	-0.303***	-0.325***
		(0.0409)	(0.0412)	(0.0416)
Native born: mother foreign, father native		-0.234***	-0.269***	-0.341***
		(0.0594)	(0.0596)	(0.0603)
Native born: both parents foreign		0.0916***	0.0537*	0.0303
		(0.0229)	(0.0234)	(0.0236)
Foreign-born		0.254***	0.225***	0.364***
		(0.0231)	(0.0236)	(0.0239)
Race				
White			-	-
Black			-0.515***	-0.476***
			(0.0592)	(0.0596)
Chinese			-1.748***	-1.693***
			(0.219)	(0.220)
Japanese			-0.568	-0.585
			(0.578)	(0.578)
Age				-0.0326***
				(0.000763)
N	152589	152589	152589	152589

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 113: Neighborhood-level Mean Income Differences between Leavers and Stayers, 1930-1940



Note: Blue shades mean leavers' mean occupational income was lower than stayers, whereas red shades mean leavers' mean occupational income was higher than stayers in the Year 1930.

Table 111: Logit Results between Leavers and Stayers at the City level: 1930-1940 males

[City Leavers]				
Occupational income	0.00198	-0.00555**	-0.00471*	-0.00676**
	(0.00194)	(0.00206)	(0.00209)	(0.00213)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-0.744***	-0.767***	-0.761***
		(0.129)	(0.129)	(0.129)
Native born: mother foreign, father native		-0.385*	-0.400**	-0.384*
		(0.152)	(0.152)	(0.152)
Native born: both parents foreign		-1.264***	-1.272***	-1.220***
		(0.0634)	(0.0642)	(0.0649)
Foreign-born		-0.681***	-0.722***	-0.778***
		(0.0555)	(0.0566)	(0.0578)
Race				
White			-	-
Black			-0.203	-0.227
			(0.177)	(0.178)
Chinese			0.751***	0.744***
			(0.158)	(0.158)
Japanese			0.935	0.836
			(0.867)	(0.868)
Age				0.0113***
				(0.00211)
[City Stayers]	(base outcome)			
N	8789	8789	8789	8789

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 112: Multinomial Logit Results between Leavers and Stayers at Neighborhood Level: 1930-1940 males

[City Core Stayers]: baseline comparison group

[City leavers & NYC metro area stayers]				
Occupational income	0.00650	-0.00143	-0.00366	0.000959
	(0.00388)	(0.00405)	(0.00410)	(0.00419)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-0.551*	-0.551*	-0.557*
		(0.258)	(0.259)	(0.260)
Native born: mother foreign, father native		0.108	0.102	0.0665
		(0.308)	(0.308)	(0.309)
Native born: both parents foreign		-0.743***	-0.762***	-0.878***
		(0.128)	(0.129)	(0.131)
Foreign-born		-0.964***	-0.924***	-0.803***
		(0.121)	(0.122)	(0.124)
Race				
White			-	-
Black			-0.391	-0.339
			(0.465)	(0.466)
Chinese			-1.594***	-1.590***
			(0.475)	(0.475)
Japanese			-13.46	-12.95
			(1221.9)	(1047.5)
Age				-0.0251***
				(0.00439)
[City & NYC metro area Leavers]				
Occupational income	-0.00398	-0.0130***	-0.0130***	-0.0110***
	(0.00295)	(0.00307)	(0.00308)	(0.00314)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-0.903***	-0.907***	-0.905***
		(0.199)	(0.199)	(0.200)
Native born: mother foreign, father native		-0.439	-0.433	-0.449
		(0.260)	(0.260)	(0.260)
Native born: both parents foreign		-1.435***	-1.429***	-1.480***
		(0.101)	(0.102)	(0.104)
Foreign-born		-0.874***	-0.865***	-0.815***
		(0.0899)	(0.0918)	(0.0932)
Race				
White			-	-
Black			0.136	0.156
			(0.322)	(0.323)
Chinese			-0.152	-0.154
			(0.179)	(0.179)
Japanese			0.461	0.546
			(1.122)	(1.123)
Age				-0.0109***
				(0.00312)

Continued

Table 112: Multinomial Logit Results between Leavers and Stayers at Neighborhood Level: 1930-1940 males (cont.)

[City Core Stayers]: baseline comparison group

[City Stayer & City Core Leavers]				
Occupational income	-0.00516	-0.00680*	-0.00875**	-0.00242
	(0.00291)	(0.00296)	(0.00298)	(0.00306)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-0.114	-0.0854	-0.0989
		(0.192)	(0.193)	(0.195)
Native born: mother foreign, father native		0.0877	0.115	0.0679
		(0.260)	(0.260)	(0.262)
Native born: both parents foreign		-0.0144	-0.00155	-0.163
		(0.0978)	(0.0990)	(0.101)
Foreign-born		-0.281**	-0.206*	-0.0365
		(0.0924)	(0.0940)	(0.0959)
Race				
White			-	-
Black			0.335	0.408
			(0.326)	(0.328)
Chinese			-2.080***	-2.070***
			(0.279)	(0.280)
Japanese			-1.029	-0.730
			(1.416)	(1.419)
Age				-0.0350***
				(0.00312)
N	8789	8789	8789	8789

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

1.4.3 Were the People Who Moved Into the Periphery Richer than Original Residents of the Periphery?

Jackson (1985) discusses Brooklyn's transformation from being essentially agricultural to the favorite residence of gentlemen of taste and fortune between the 1810s and the 1850s due to the regular steam ferry service to the NYC. During the early nineteenth century, Brooklyn became the "transit-hub" connected to the center of the city, and the influx of middle-class families changed the orientation of neighborhoods — "the little village of Bedford (now part of Bedford-Stuyvesant in Northeast Brooklyn), for example, used to be essentially rural until 1850. However, after the influx of middle-class families, and it had become part of the expanding metropolis, very few laborers remained, and the farmers had disappeared."[13]

If the pattern of early nineteenth-century Brooklyn as the periphery of the city as the transit spoke held for my study period, the longitudinal database should reveal that entrants who moved into the periphery were richer than original residents of the periphery of the city. Hence, I take the longitudinal data of individuals and compare the mean occupational income of residents who moved into the periphery and who stayed in the periphery at the NTA level.

The longitudinal data reveals that the entrants who moved into the periphery were *not* richer than the original residents of the periphery. For example, Figures 115, 116, 117 show that the entrants to the periphery had mostly lower mean occupational income than the original residents. Given each NTA in the City, I take the difference of mean occupational income of entrants and stayers at the periphery over the study period. In Figures 114, 115, 116, 117, given each NTA, blue shades (the darker blue, the poorer entrants) indicate entrants being poorer, whereas red shades (the darker red, the richer entrants) indicate entrants being richer than the stayers. Data during the study period indicates that the entrants to the periphery were, in fact, *not* richer than the stayers.

Regression results also support that new suburbanites were not richer than the people who already lived at the periphery. I run logit regression to model the log odds of individuals' entering into the city periphery, using the longitudinal data of individuals during the study period. The predictor variables of interest are occupational income, nativity, race, and age. Regression results in Tables 113, 115, 117, 119 show that as one's occupational income increases, the log odds of moving into the city periphery decreases. In terms of the nativity, being foreign-born relative to native-born with both native parents (which may be associated with one's "prestige to the public's eye) decreases the log odds of entering the city periphery. In terms of race, being non-white relative to white increases the log odds of moving into the periphery and the degree of log odds of outcome varies across race; however, considering that the majority of residents in New York were white, this may need to be interpreted with

[13] Recited from Jackson (1985), originally from Gilman (1971).

caution. Finally, older people were more likely to enter into the city periphery throughout the study period.

While the regression results in Tables 113, 115, 117, 119 look at extensive margin of entering to the periphery of the city regardless of the nature of flows (i.e. whether entrants migrated to the city's periphery from NYC metro area, or migrated from Alabama, or migrated from the core of the city), Tables 114, 116, 118, 120, 122 look at whether flows from the metro area to the city periphery may have been different from flows from outside the metro area to the city periphery, as well as flows from the city core to the city periphery. Regression results regarding city periphery entrants that separately looks at entrants with varying origins tells us consistent story (as in periphery entrants at the extensive margin regardless of origins) that relative to people who stayed in the city periphery, as occupational income increases, people who lived somewhere other than the city periphery (at any migration origins ranging from the city core to outside NYC metro area) were less likely to migrate to the city periphery. In terms of race, being non-white relative to white has varying degree and signs depending on origins of migration, however, considering that the majority of residents in New York were white, this may need to be interpreted with caution. Finally, older people were less likely to migrate into the city periphery up until the Great Depression.

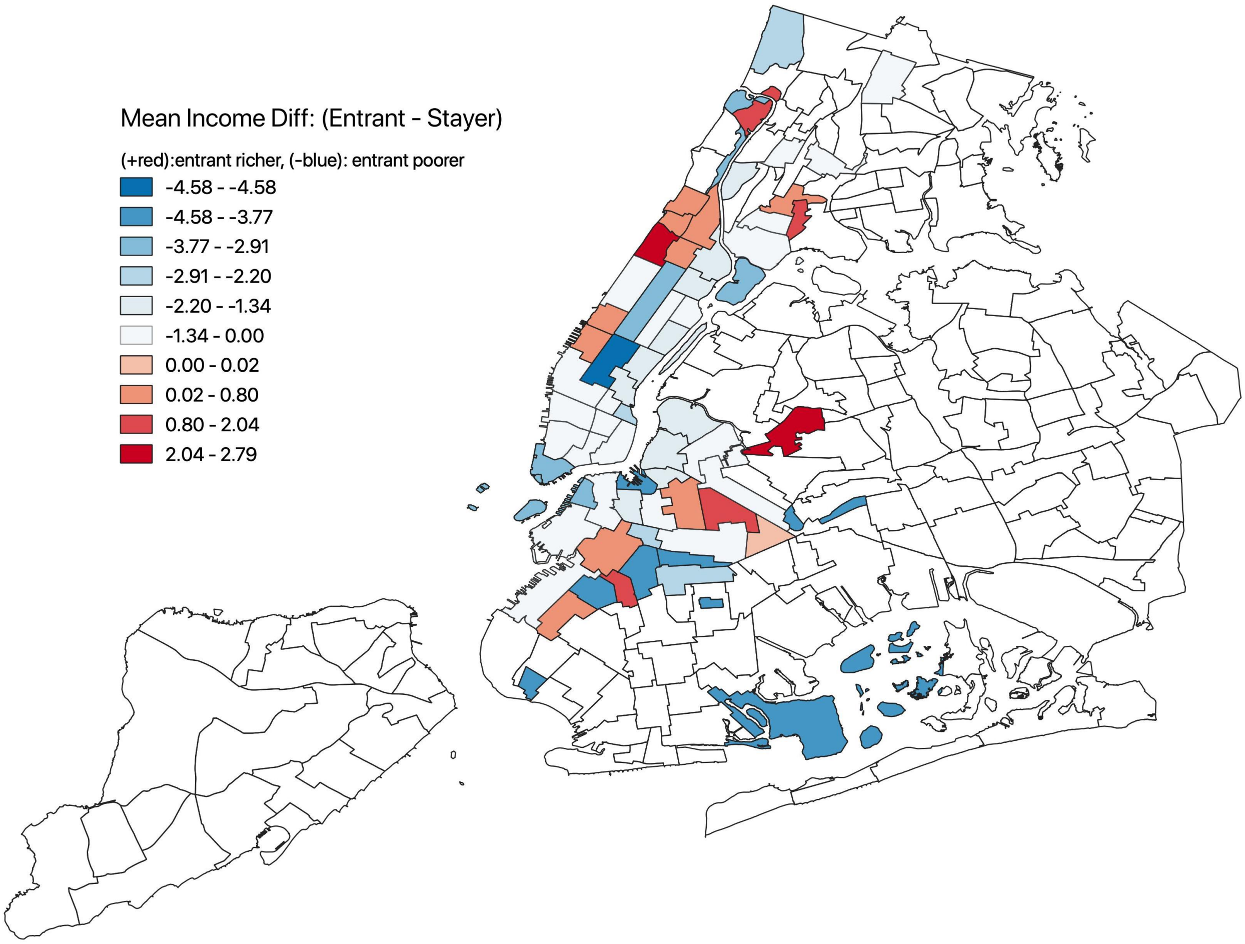
Related to my earlier discussion of income measures feature of reflecting occupation only and not reflecting other factors such as nativity and age that may have determined one's income should be noted in interpreting suburbanites' pattern of migration.[14] To the public's eyes, migration of the older to the periphery may have been associated as the movement of the "affluent." However, this only makes the accuracy of income measures more crucial.

Section 1.4.4.2 discusses decomposition of various flows of the periphery of the city including the relative income difference between entrants and stayers as well as the corresponding relative magnitudes of those flows at the neighborhood level. Regarding the relative income

[14] Features of occupational income measures are discussed in Subsection 1.4.2.

growth at the periphery, periphery entrants from anywhere (from the city core, NYC metro area, and outside NYC metro area) had higher mean income than the periphery leaving NYC metro area at all, and the relative magnitude of inflows were much bigger than outflows, making the periphery income increase. Furthermore, as Figure 122b shows, people who stayed at the periphery got richer as the metropolis grew.

Figure 114: Neighborhood-level Mean Income Differences between Entrants and Stayers, 1880-1900



Note: Blue shades mean entrants' mean occupational income was lower than stayers, whereas red shades mean entrants' mean occupational income was higher than stayers in 1900.

Table 113: Logit Results between Entrants and Stayers at the City level: 1880-1900 males

[City Entrants]				
Occupational income	-0.00169	-0.00674***	-0.00538**	-0.00554**
	(0.00160)	(0.00171)	(0.00174)	(0.00175)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-1.529***	-1.514***	-1.548***
		(0.153)	(0.153)	(0.153)
Native born: mother foreign, father native		-1.125***	-1.119***	-1.133***
		(0.221)	(0.221)	(0.221)
Native born: both parents foreign		-1.832***	-1.812***	-1.899***
		(0.0542)	(0.0546)	(0.0563)
Foreign-born		-0.650***	-0.645***	-0.603***
		(0.0425)	(0.0432)	(0.0436)
Race				
White			-	-
Black			0.390*	0.330*
			(0.154)	(0.154)
Chinese			2.578***	2.484***
			(0.594)	(0.594)
Age				-0.0134***
				(0.00196)
[City Stayers]	(base outcome)			
N	13239	13239	13239	13239

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 114: Multinomial Logit Results between Entrants and Stayers at Neighborhood Level: 1880-1900 males

[City Periphery Stayers]: baseline comparison group				
[Periphery Entrants from NYC metro area]				
Occupational income	0.00473	-0.00355	-0.00357	-0.00414
	(0.00506)	(0.00519)	(0.00529)	(0.00531)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-1.431**	-1.431**	-1.589***
		(0.474)	(0.474)	(0.475)
Native born: mother foreign, father native		-0.399	-0.399	-0.453
		(0.594)	(0.594)	(0.596)
Native born: both parents foreign		-1.601***	-1.601***	-1.907***
		(0.166)	(0.167)	(0.174)
Foreign-born		-1.209***	-1.214***	-1.084***
		(0.147)	(0.148)	(0.150)
Race				
White			-	-
Black			0.0590	-0.139
			(0.613)	(0.614)
Chinese			12.67	13.69
			(452.3)	(887.0)
Age				-0.0432***
				(0.00620)
[Periphery Entrants from Non-NYC metro area]				
Occupational income	-0.00437	-0.0106**	-0.00881*	-0.00933*
	(0.00375)	(0.00391)	(0.00397)	(0.00399)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-1.737***	-1.713***	-1.860***
		(0.306)	(0.307)	(0.309)
Native born: mother foreign, father native		-1.379**	-1.368**	-1.416**
		(0.487)	(0.488)	(0.490)
Native born: both parents foreign		-2.092***	-2.061***	-2.346***
		(0.118)	(0.119)	(0.125)
Foreign-born		-0.828***	-0.812***	-0.693***
		(0.110)	(0.111)	(0.112)
Race				
White			-	-
Black			0.730	0.546
			(0.514)	(0.515)
Chinese			13.78	14.82
			(452.3)	(887.0)
Age				-0.0396***
				(0.00439)

Continued

Table 114: Multinomial Logit Results between Entrants and Stayers at Neighborhood Level: 1880-1900 males (cont.)

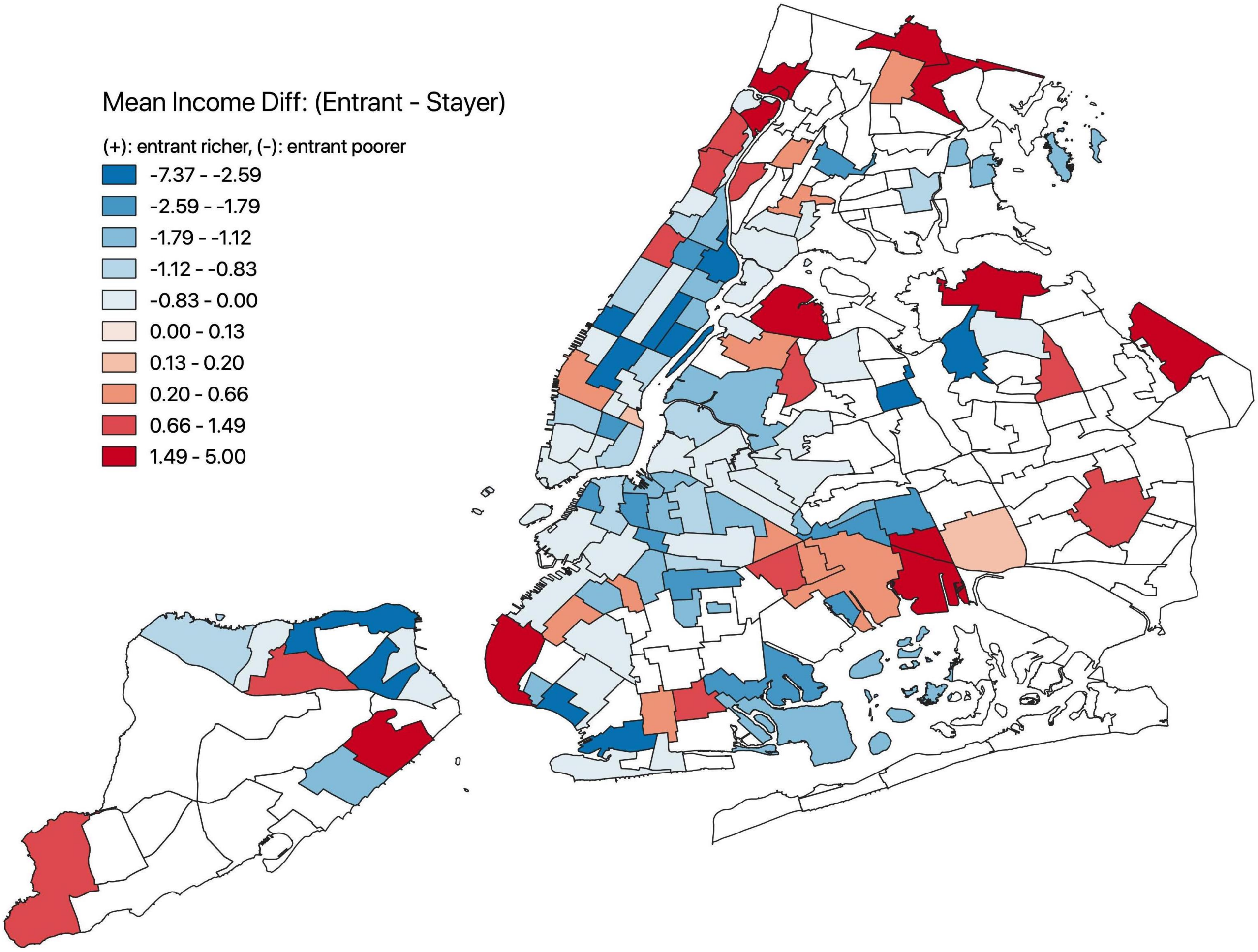
[City Periphery Stayers]: baseline comparison group

[Periphery Entrants from the City's Core]				
Occupational income	-0.00226	-0.00367	-0.00340	-0.00379
	(0.00383)	(0.00395)	(0.00401)	(0.00402)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-0.205	-0.196	-0.313
		(0.298)	(0.298)	(0.299)
Native born: mother foreign, father native		-0.154	-0.149	-0.188
		(0.483)	(0.483)	(0.484)
Native born: both parents foreign		-0.242*	-0.232*	-0.457***
		(0.117)	(0.118)	(0.123)
Foreign-born		-0.225*	-0.215	-0.126
		(0.113)	(0.114)	(0.115)
Race				
White			-	-
Black			0.331	0.189
			(0.528)	(0.529)
Chinese			11.29	12.40
			(452.3)	(887.0)
Age				-0.0299***
				(0.00446)
N	13239	13239	13239	13239

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 115: Neighborhood-level Mean Income Differences between Entrants and Stayers, 1900-1910



Note: Blue shades mean entrants' mean occupational income was lower than stayers, whereas red shades mean entrants' mean occupational income was higher than stayers in 1910.

Table 115: Logit Results between Entrants and Stayers: 1900-1910 males

[City Entrants]				
Occupational income	-0.00774***	-0.00671***	-0.00608***	-0.00671***
	(0.000945)	(0.000986)	(0.000998)	(0.00100)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-1.354***	-1.328***	-1.272***
		(0.0710)	(0.0713)	(0.0716)
Native born: mother foreign, father native		-1.207***	-1.181***	-1.148***
		(0.114)	(0.114)	(0.114)
Native born: both parents foreign		-1.189***	-1.161***	-1.162***
		(0.0330)	(0.0338)	(0.0338)
Foreign-born		-0.143***	-0.117***	-0.170***
		(0.0287)	(0.0297)	(0.0301)
Race				
White			-	-
Black			0.331***	0.367***
			(0.0940)	(0.0941)
Chinese			1.360**	1.365**
			(0.427)	(0.428)
Japanese			14.06	14.12
			(1265.8)	(1265.7)
Age				0.0118***
				(0.00101)
[City Stayers]	(base outcome)			
N	47201	47201	47201	47201

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 116: Multinomial Logit Results between Entrants and Stayers at Neighborhood Level: 1900-1910 males

[City Periphery Stayers]: baseline comparison group				
[Periphery Entrants from NYC metro area]				
Occupational income	-0.0121***	-0.0109***	-0.0112***	-0.0100***
	(0.00256)	(0.00261)	(0.00265)	(0.00266)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-1.019***	-1.025***	-1.144***
		(0.173)	(0.174)	(0.175)
Native born: mother foreign, father native		-1.149***	-1.156***	-1.227***
		(0.294)	(0.295)	(0.295)
Native born: both parents foreign		-0.463***	-0.471***	-0.479***
		(0.0848)	(0.0863)	(0.0864)
Foreign-born		0.0536	0.0453	0.144
		(0.0760)	(0.0778)	(0.0787)
Race				
White			-	-
Black			-0.0542	-0.136
			(0.278)	(0.278)
Chinese			0.339	0.310
			(1.226)	(1.226)
Japanese			-0.430	-0.584
			(2134.9)	(3520.6)
Age				-0.0224***
				(0.00266)
[Periphery Entrants from Non-NYC metro area]				
Occupational income	-0.0145***	-0.0118***	-0.0108***	-0.0100***
	(0.00186)	(0.00191)	(0.00194)	(0.00195)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-1.616***	-1.574***	-1.658***
		(0.115)	(0.115)	(0.116)
Native born: mother foreign, father native		-1.474***	-1.431***	-1.481***
		(0.181)	(0.181)	(0.181)
Native born: both parents foreign		-1.048***	-1.002***	-1.011***
		(0.0638)	(0.0650)	(0.0651)
Foreign-born		0.181**	0.225***	0.292***
		(0.0589)	(0.0603)	(0.0609)
Race				
White			-	-
Black			0.609**	0.550*
			(0.214)	(0.215)
Chinese			1.327	1.298
			(1.009)	(1.009)
Japanese			12.58	13.47
			(1555.1)	(2564.6)
Age				-0.0154***
				(0.00199)

Continued

Table 116: Multinomial Logit Results between Entrants and Stayers at Neighborhood Level: 1900-1910 males (cont.)

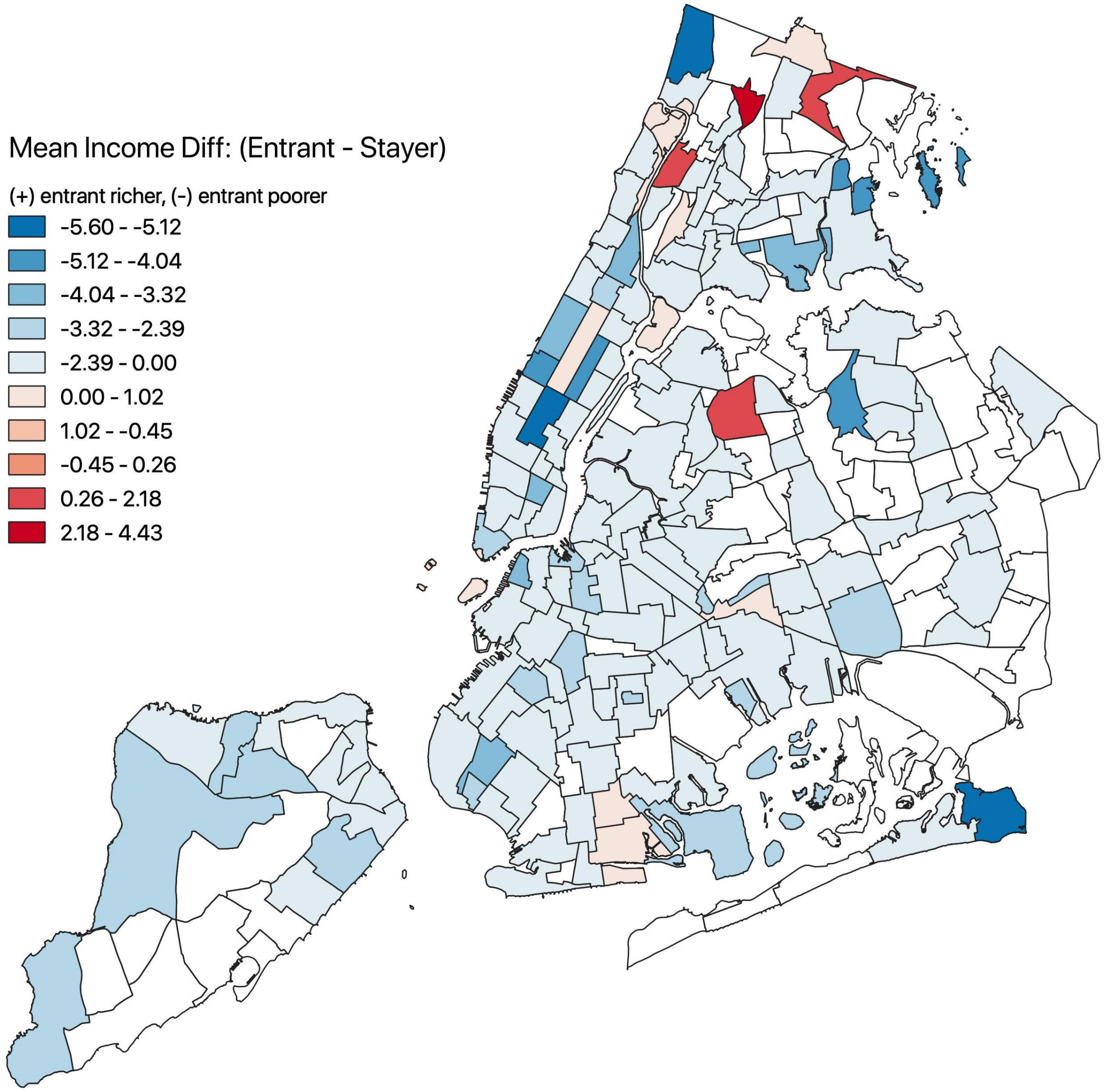
`[City Periphery Stayers]:  baseline comparison group`

[Periphery Entrants from the City's Core]				
Occupational income	-0.00803***	-0.00620**	-0.00585**	-0.00412*
	(0.00198)	(0.00203)	(0.00206)	(0.00208)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-0.270*	-0.254*	-0.426***
		(0.118)	(0.119)	(0.120)
Native born: mother foreign, father native		-0.317	-0.301	-0.405*
		(0.189)	(0.189)	(0.191)
Native born: both parents foreign		0.246***	0.264***	0.260***
		(0.0679)	(0.0692)	(0.0695)
Foreign-born		0.387***	0.405***	0.560***
		(0.0640)	(0.0655)	(0.0663)
Race				
White			-	-
Black			0.297	0.184
			(0.231)	(0.231)
Chinese			-0.0966	-0.133
			(1.096)	(1.096)
Japanese			-0.272	-0.486
			(1715.0)	(2828.2)
Age				-0.0344***
				(0.00214)
N	47201	47201	47201	47201

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 116: Neighborhood-level Mean Income Differences between Entrants and Stayers, 1910-1920



Note: Blue shades mean entrants' mean occupational income was lower than stayers, whereas red shades mean entrants' mean occupational income was higher than stayers in 1920.

Table 117: Logit Results between Entrants and Stayers at the City level: 1910-1920 males

[Entrants]				
Occupational income	-0.0166***	-0.0147***	-0.0134***	-0.0135***
	(0.000431)	(0.000443)	(0.000448)	(0.000448)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-1.618***	-1.564***	-1.555***
		(0.0345)	(0.0346)	(0.0346)
Native born: mother foreign, father native		-1.551***	-1.498***	-1.492***
		(0.0541)	(0.0542)	(0.0542)
Native born: both parents foreign		-1.205***	-1.147***	-1.151***
		(0.0149)	(0.0152)	(0.0152)
Foreign-born		-0.115***	-0.0603***	-0.0737***
		(0.0114)	(0.0118)	(0.0119)
Race				
White			-	-
Black			0.568***	0.572***
			(0.0328)	(0.0328)
Chinese			1.556***	1.545***
			(0.222)	(0.222)
Japanese			2.864***	2.889***
			(0.591)	(0.591)
Age				0.00385***
				(0.000426)
[Stayers]	(base outcome)			
N	206052	206052	206052	206052

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 118: Multinomial Logit Results between Entrants and Stayers at Neighborhood Level: 1910-1920 males

[City Periphery Stayers]: baseline comparison group				
[Periphery Entrants from NYC metro area]				
Occupational income	-0.0177***	-0.0154***	-0.0160***	-0.0154***
	(0.00108)	(0.00109)	(0.00111)	(0.00111)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-1.075***	-1.100***	-1.161***
		(0.0724)	(0.0727)	(0.0728)
Native born: mother foreign, father native		-1.155***	-1.179***	-1.222***
		(0.119)	(0.119)	(0.119)
Native born: both parents foreign		-0.636***	-0.663***	-0.641***
		(0.0346)	(0.0352)	(0.0353)
Foreign-born		0.205***	0.176***	0.252***
		(0.0287)	(0.0295)	(0.0298)
Race				
White			-	-
Black			-0.221*	-0.255**
			(0.0907)	(0.0908)
Chinese			2.146	2.191*
			(1.096)	(1.096)
Japanese			14.34	13.66
			(1072.8)	(832.2)
Age				-0.0224***
				(0.00105)
[Periphery Entrants from Non-NYC metro area]				
Occupational income	-0.0231***	-0.0193***	-0.0178***	-0.0172***
	(0.000652)	(0.000673)	(0.000681)	(0.000683)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-1.888***	-1.833***	-1.889***
		(0.0451)	(0.0453)	(0.0454)
Native born: mother foreign, father native		-1.796***	-1.743***	-1.783***
		(0.0697)	(0.0698)	(0.0701)
Native born: both parents foreign		-1.246***	-1.188***	-1.168***
		(0.0220)	(0.0225)	(0.0226)
Foreign-born		0.272***	0.327***	0.396***
		(0.0188)	(0.0194)	(0.0196)
Race				
White			-	-
Black			0.545***	0.513***
			(0.0551)	(0.0553)
Chinese			3.225**	3.264**
			(1.005)	(1.006)
Japanese			16.31	15.65
			(1072.8)	(832.2)
Age				-0.0204***
				(0.000664)

Continued

Table 118: Multinomial Logit Results between Entrants and Stayers at Neighborhood Level: 1910-1920 males (cont.)

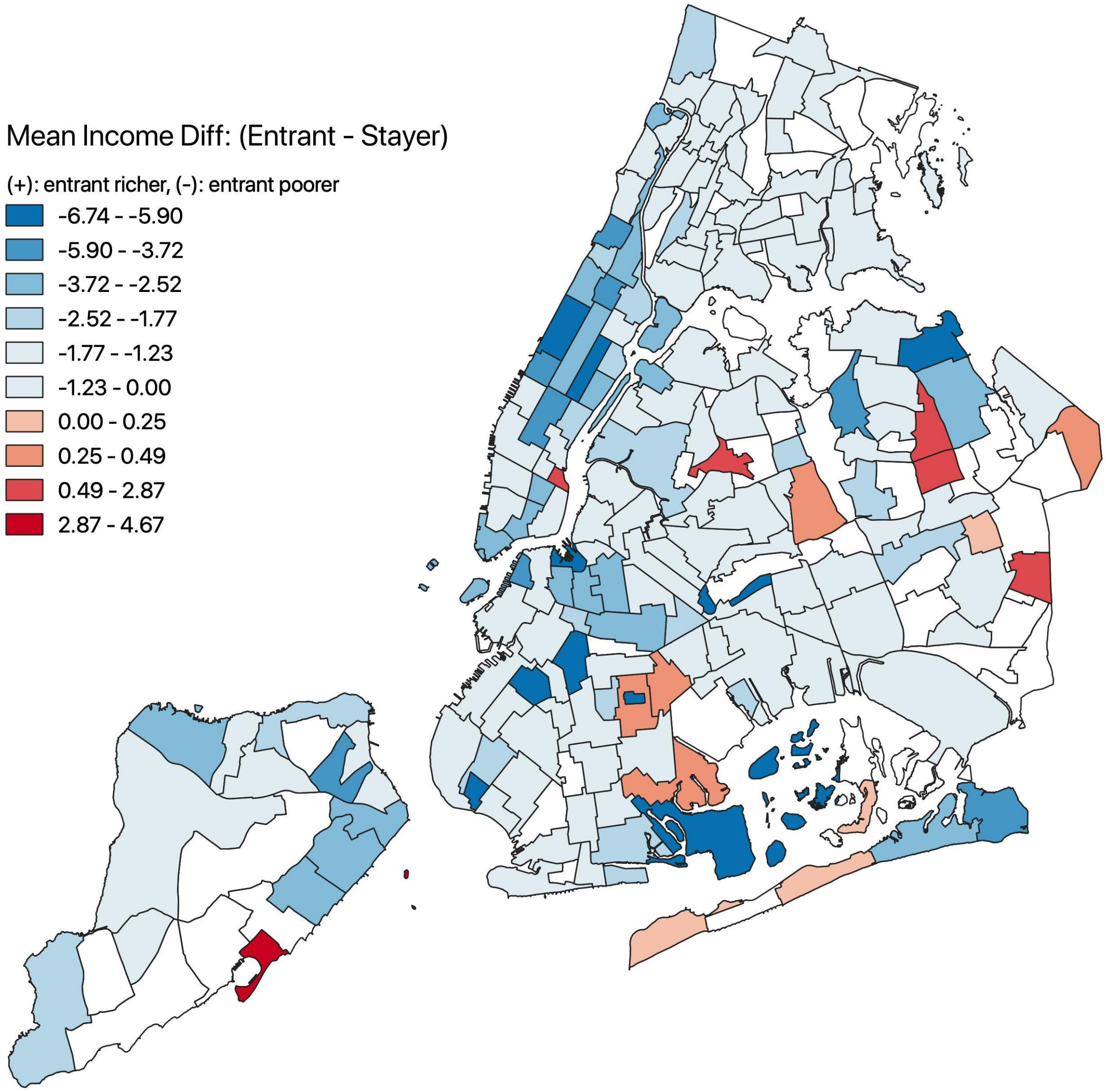
[City Periphery Stayers]: baseline comparison group

[Periphery Entrants from the City's Core]				
Occupational income	-0.00775***	-0.00536***	-0.00549***	-0.00462***
	(0.000629)	(0.000644)	(0.000651)	(0.000656)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-0.198***	-0.210***	-0.293***
		(0.0351)	(0.0353)	(0.0357)
Native born: mother foreign, father native		-0.213***	-0.224***	-0.284***
		(0.0544)	(0.0546)	(0.0552)
Native born: both parents foreign		0.0650**	0.0524*	0.0838***
		(0.0209)	(0.0213)	(0.0215)
Foreign-born		0.492***	0.479***	0.591***
		(0.0195)	(0.0200)	(0.0203)
Race				
White			-	-
Black			-0.115	-0.160**
			(0.0605)	(0.0608)
Chinese			1.764	1.837
			(1.020)	(1.021)
Japanese			13.55	12.82
			(1072.8)	(832.2)
Age				-0.0321***
				(0.000660)
N	206052	206052	206052	206052

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 117: Neighborhood-level Mean Income Differences between Entrants and Stayers, 1920-1930



Note: Blue shades mean entrants' mean occupational income was lower than stayers, whereas red shades mean entrants' mean occupational income was higher than stayers in 1930.

Table 119: Logit Results between Entrants and Stayers at the City level: 1920-1930 males

[City Entrants]				
Occupational income	-0.0210***	-0.0152***	-0.0119***	-0.0119***
	(0.000361)	(0.000375)	(0.000379)	(0.000379)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-1.877***	-1.761***	-1.761***
		(0.0334)	(0.0335)	(0.0335)
Native born: mother foreign, father native		-1.750***	-1.632***	-1.631***
		(0.0473)	(0.0474)	(0.0474)
Native born: both parents foreign		-1.611***	-1.486***	-1.486***
		(0.0124)	(0.0128)	(0.0128)
Foreign-born		0.323***	0.437***	0.433***
		(0.00867)	(0.00907)	(0.00924)
Race				
White			-	-
Black			0.942***	0.943***
			(0.0192)	(0.0193)
Chinese			1.780***	1.781***
			(0.187)	(0.187)
Japanese			1.623***	1.626***
			(0.307)	(0.307)
Age				0.000872*
				(0.000358)
[City Stayers]	(base outcome)			
N	363947	363947	363947	363947

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 120: Multinomial Logit Results between Entrants and Stayers at Neighborhood Level: 1920-1930 males

[City Periphery Stayers]: baseline comparison group				
[Periphery Entrants from NYC metro area]				
Occupational income	-0.0149***	-0.00873***	-0.00984***	-0.00936***
	(0.000924)	(0.000925)	(0.000939)	(0.000940)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-1.207***	-1.268***	-1.276***
		(0.0704)	(0.0706)	(0.0706)
Native born: mother foreign, father native		-1.193***	-1.255***	-1.299***
		(0.106)	(0.106)	(0.106)
Native born: both parents foreign		-0.677***	-0.741***	-0.758***
		(0.0300)	(0.0305)	(0.0305)
Foreign-born		0.891***	0.831***	0.936***
		(0.0241)	(0.0246)	(0.0251)
Race				
White			-	-
Black			-0.647***	-0.685***
			(0.0643)	(0.0644)
Chinese			2.590*	2.591*
			(1.098)	(1.098)
Japanese			-0.534	-0.641
			(1.155)	(1.155)
Age				-0.0218***
				(0.000883)
[Periphery Entrants from Non-NYC metro area]				
Occupational income	-0.0227***	-0.0155***	-0.0127***	-0.0122***
	(0.000506)	(0.000524)	(0.000529)	(0.000531)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-2.253***	-2.169***	-2.177***
		(0.0406)	(0.0408)	(0.0409)
Native born: mother foreign, father native		-2.049***	-1.964***	-2.013***
		(0.0568)	(0.0570)	(0.0571)
Native born: both parents foreign		-1.609***	-1.519***	-1.539***
		(0.0166)	(0.0170)	(0.0171)
Foreign-born		0.644***	0.726***	0.843***
		(0.0135)	(0.0141)	(0.0144)
Race				
White			-	-
Mexican			1.943**	1.762*
			(0.724)	(0.724)
Black			0.526***	0.483***
			(0.0273)	(0.0275)
Chinese			4.023***	4.023***
			(1.007)	(1.008)
Japanese			1.513*	1.395*
			(0.592)	(0.593)
Age				-0.0243***
				(0.000502)

Continued

Table 120: Multinomial Logit Results between Entrants and Stayers at Neighborhood Level: 1920-1930 males (cont.)

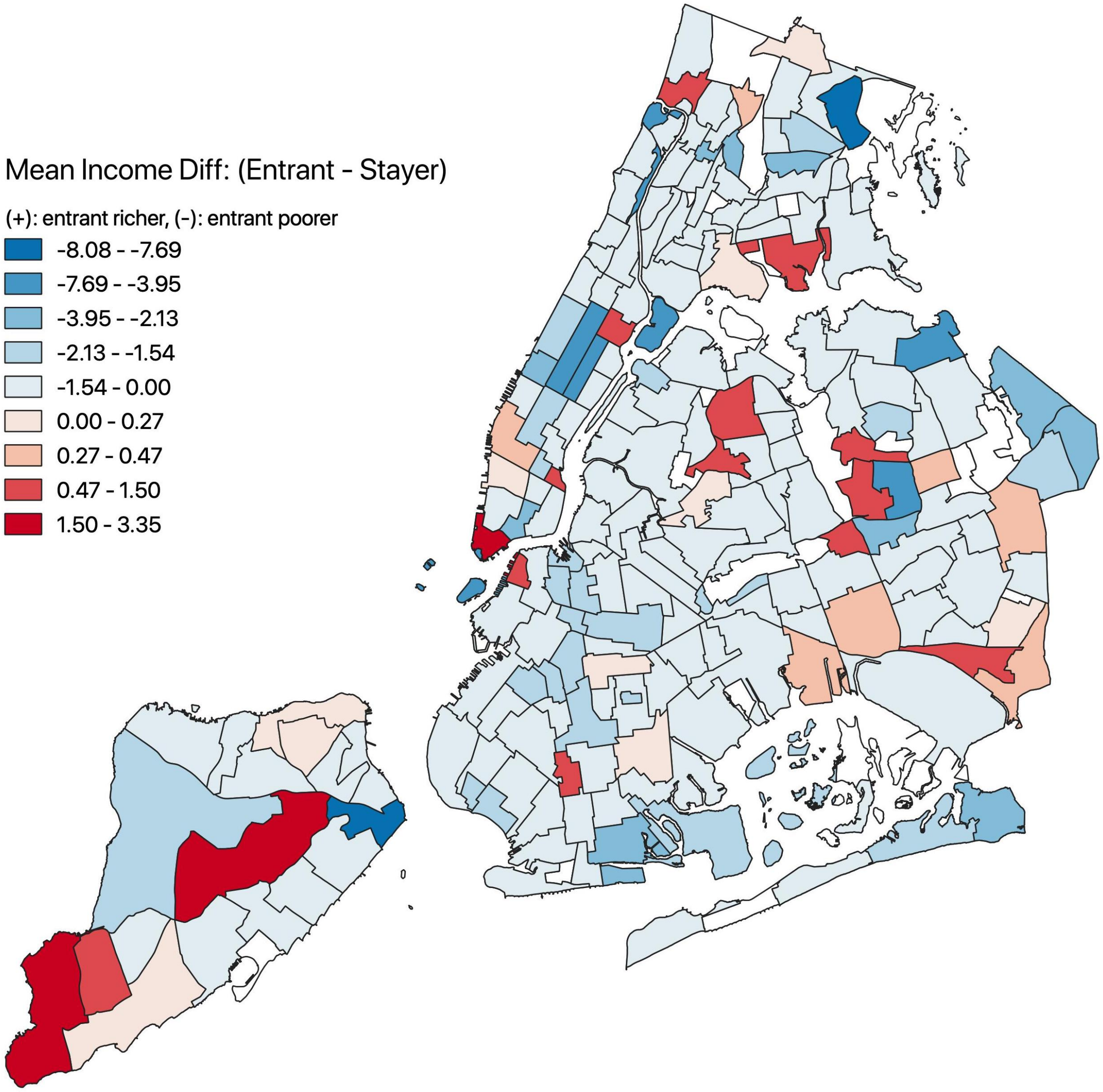
`[City Periphery Stayers]:  baseline comparison group`

[Periphery Entrants from the City's Core]				
Occupational income	-0.000953*	0.000674	-0.000479	0.000176
	(0.000434)	(0.000443)	(0.000448)	(0.000452)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-0.307***	-0.368***	-0.376***
		(0.0226)	(0.0228)	(0.0230)
Native born: mother foreign, father native		-0.254***	-0.316***	-0.380***
		(0.0328)	(0.0330)	(0.0334)
Native born: both parents foreign		0.181***	0.116***	0.0883***
		(0.0128)	(0.0132)	(0.0133)
Foreign-born		0.447***	0.387***	0.545***
		(0.0132)	(0.0135)	(0.0138)
Race				
White			-	-
Black			-0.715***	-0.771***
			(0.0304)	(0.0306)
Chinese			2.330*	2.326*
			(1.013)	(1.013)
Japanese			-0.274	-0.424
			(0.658)	(0.659)
Age				-0.0325***
				(0.000441)
N	363947	363947	363947	363947

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Figure 118: Neighborhood-level Mean Income Differences between Entrants and Stayers, 1930-1940



Note: Blue shades mean entrants' mean occupational income was lower than stayers, whereas red shades mean entrants' mean occupational income was higher than stayers in 1940.

Table 121: Logit Results between Entrants and Stayers at the City level: 1930-1940 males

[City Entrants]				
Occupational income	-0.0111***	-0.00901***	-0.00757***	-0.00792***
	(0.00110)	(0.00114)	(0.00116)	(0.00116)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-1.285***	-1.216***	-1.223***
		(0.0875)	(0.0879)	(0.0882)
Native born: mother foreign, father native		-1.424***	-1.356***	-1.364***
		(0.123)	(0.123)	(0.123)
Native born: both parents foreign		-1.471***	-1.400***	-1.360***
		(0.0342)	(0.0352)	(0.0354)
Foreign-born		-0.339***	-0.271***	-0.374***
		(0.0278)	(0.0289)	(0.0300)
Race				
White			-	-
Black			0.483***	0.525***
			(0.0607)	(0.0609)
Chinese			0.811	0.898
			(0.524)	(0.523)
Japanese			-14.13	-13.86
			(693.0)	(611.5)
Age				0.0163***
				(0.00117)
[City Stayers]	(base outcome)			
N	169642	169642	169642	169642

Standard errors in parentheses

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table 122: Multinomial Logit Results between Entrants and Stayers at Neighborhood Level: 1930-1940 males

[City Periphery Stayers]: baseline comparison group				
[Periphery Entrants from NYC metro area]				
Occupational income	-0.00632*	-0.00599*	-0.00911***	-0.00891**
	(0.00265)	(0.00267)	(0.00273)	(0.00274)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-1.083***	-1.218***	-1.216***
		(0.168)	(0.169)	(0.169)
Native born: mother foreign, father native		-1.104***	-1.238***	-1.235***
		(0.243)	(0.244)	(0.244)
Native born: both parents foreign		-0.711***	-0.850***	-0.876***
		(0.0761)	(0.0784)	(0.0787)
Foreign-born		-0.633***	-0.766***	-0.704***
		(0.0690)	(0.0713)	(0.0732)
Race				
White			-	-
Black			-1.294***	-1.323***
			(0.180)	(0.180)
Chinese			-13.38	-13.43
			(1029.3)	(1028.7)
Japanese			0.110	0.0923
			(6244.5)	(6244.0)
Age				-0.0102***
				(0.00271)
[Periphery Entrants from Non-NYC metro area]				
Occupational income	-0.0122***	-0.00986***	-0.00877***	-0.00871***
	(0.00180)	(0.00185)	(0.00186)	(0.00186)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-1.992***	-1.967***	-1.967***
		(0.120)	(0.121)	(0.121)
Native born: mother foreign, father native		-1.939***	-1.914***	-1.914***
		(0.170)	(0.171)	(0.171)
Native born: both parents foreign		-1.585***	-1.558***	-1.565***
		(0.0573)	(0.0598)	(0.0600)
Foreign-born		-0.383***	-0.358***	-0.343***
		(0.0498)	(0.0524)	(0.0536)
Race				
White			-	-
Black			0.134	0.126
			(0.0981)	(0.0983)
Chinese			1.074	1.069
			(1.037)	(1.038)
Japanese			-0.180	-0.186
			(4296.4)	(4297.2)
Age				-0.00253
				(0.00188)

Continued

Table 122: Multinomial Logit Results between Entrants and Stayers at Neighborhood Level: 1930-1940 males (cont.)

[City Periphery Stayers]: baseline comparison group				
[Periphery Entrants from the City's Core]				
Occupational income	-0.000461	-0.000403	-0.00162	-0.00105
	(0.00182)	(0.00184)	(0.00186)	(0.00187)
Nativity				
Native born: both parents native		-	-	-
Native born: father foreign, mother native		-0.779***	-0.861***	-0.856***
		(0.112)	(0.114)	(0.114)
Native born: mother foreign, father native		-0.510***	-0.590***	-0.583***
		(0.153)	(0.154)	(0.155)
Native born: both parents foreign		0.0493	-0.0346	-0.0978
		(0.0564)	(0.0590)	(0.0594)
Foreign-born		-0.0910	-0.171**	-0.00649
		(0.0530)	(0.0555)	(0.0568)
Race				
White			-	-
Black			-0.623***	-0.693***
			(0.112)	(0.112)
Chinese			0.176	0.0472
			(1.119)	(1.119)
Japanese			16.06	16.03
			(3887.9)	(3889.2)
Age				-0.0263***
				(0.00194)
N	32523	32523	32523	32523

Standard errors in parentheses
* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

1.4.4 Income Changes between the Core and the Periphery

I discuss the aggregate results on relative income change between the core and periphery of the city and show how the results on the flows are compatible with the aggregate results. Section 1.4.4.1 discusses the trend of mean occupational income over the study period. Section 1.4.4.2 discusses decomposition of flows among the core, the periphery, and the rest of the country, along with associated income as well as the relative magnitudes of the flows.

The geographic units of analyses here are NTAs and I use the distance from the Battery to centroids of NTAs in the city. I define the core of the city as "transit hub" neighborhoods such as Downtown Manhattan and Midtown Manhattan where transit infrastructures were extremely highly concentrated, whereas the periphery of the city are "transit spoke" neighborhoods such as Upper Manhattan and outer boroughs of the city.

1.4.4.1 Occupational income

Figure 119 reveals that the mean occupational income decreases as the distance from the Battery increases while the slope of fitted values getting more flat from 1870 to 1900. In 1910, this slope stays flat, and then in 1920, the mean occupational income slightly increases as the distance from the Battery increases. In Years 1930 and 1940, the slope becomes even steeper, implying that the mean occupational income at the edge (relative to the center) of the city increases even further. Over time, the relative income was higher in the center of the city (relative to the periphery) in the first half of the study period, whereas the relative income in the periphery became higher in the second half of the study period.

Figure 120 shows that the percent change of mean occupation income gets higher as it gets further away from the center. However, note that in the earlier study period (till 1900), mean occupational income was increasing both at the center and the edge whereas, during 1910 and 1930, percent change of mean occupational income was negative at the center which means the mean occupational income was decreasing at the center of the city. At the very end of the study period (i.e. 1930-1940), the curve becomes close to nil both at the center and edge, implying that the mean occupational income almost stayed the same as the earlier decade.

Figure 119: Mean Occupational Income

(a) 1870

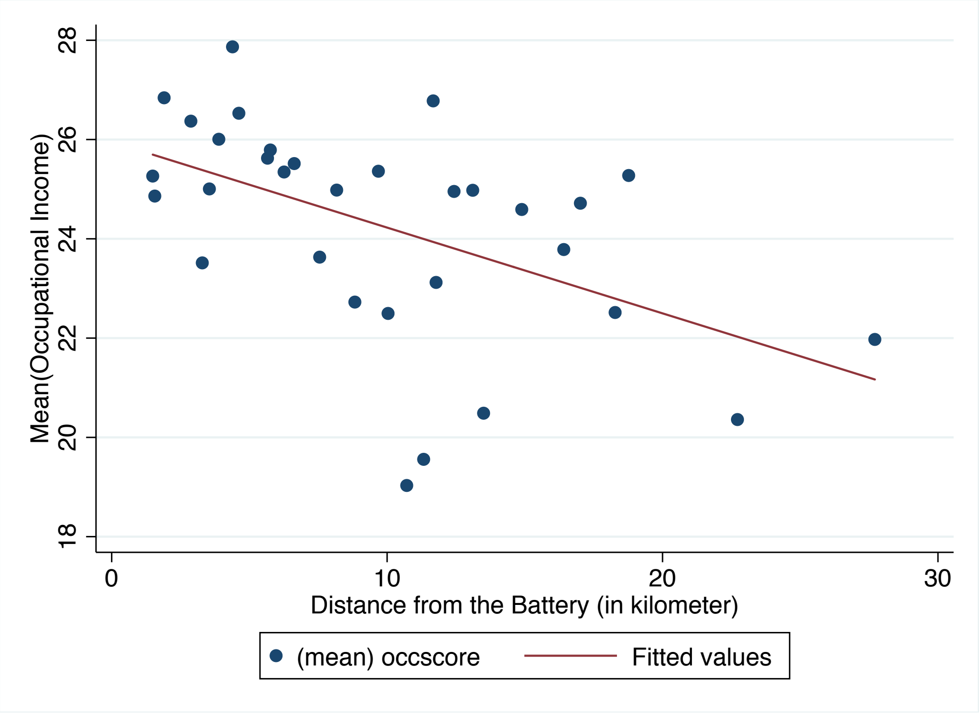
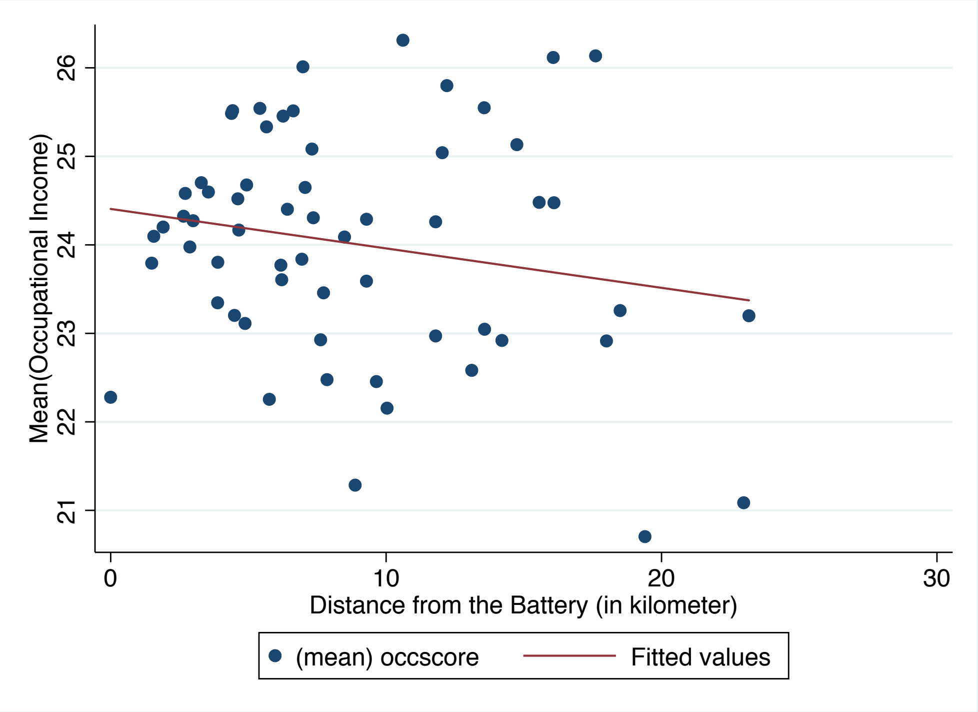


Figure 119: Mean Occupational Income

(b) 1880



(c) 1900

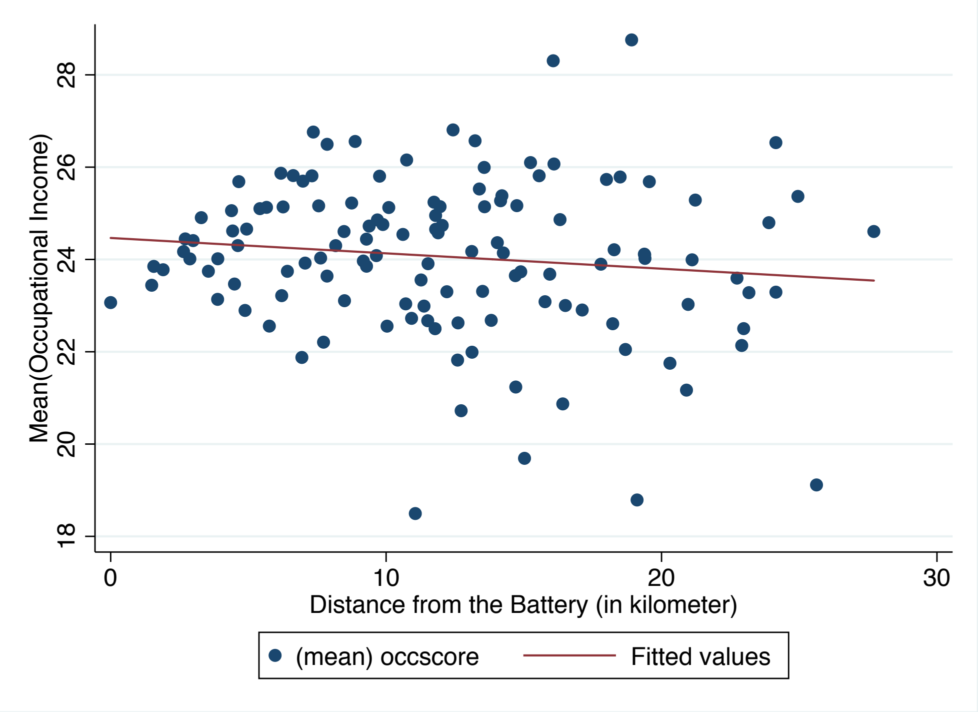
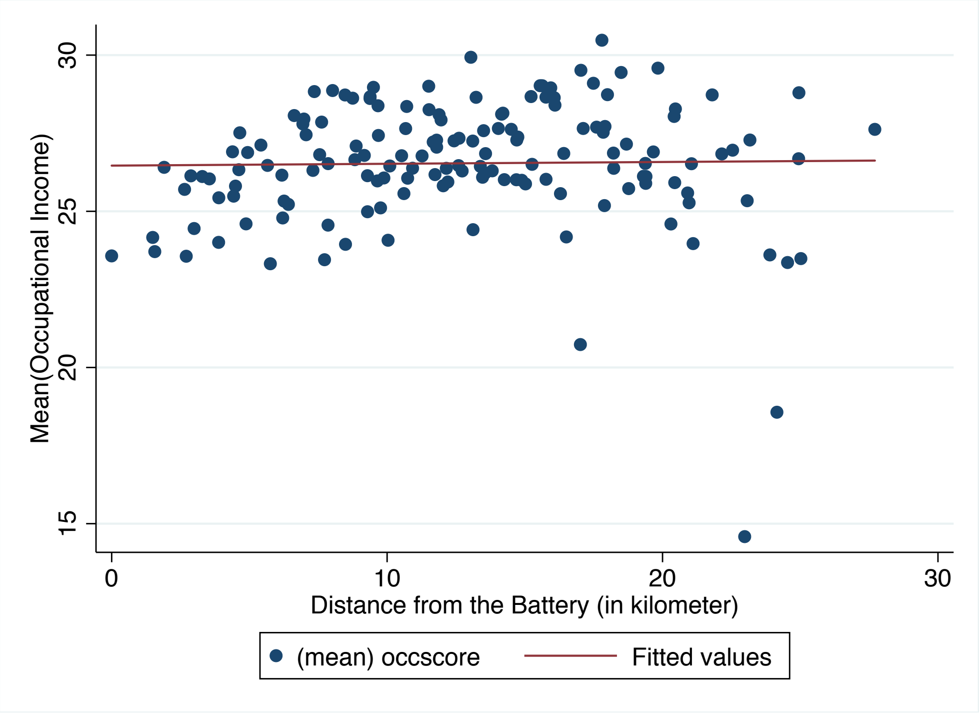


Figure 119: Mean Occupational Income

(d) 1910



(e) 1920

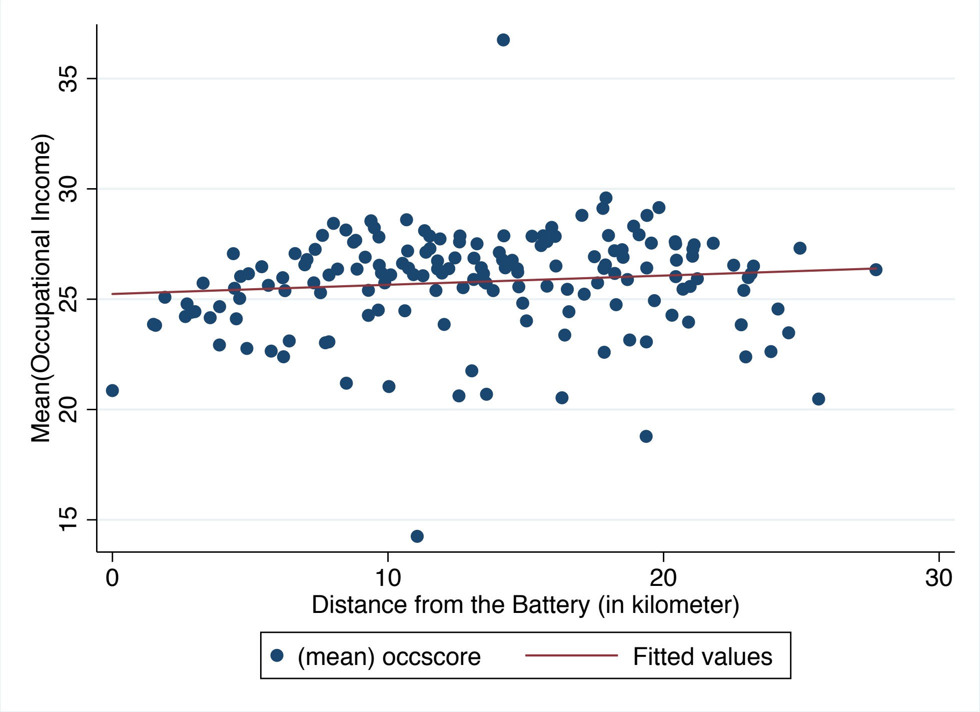
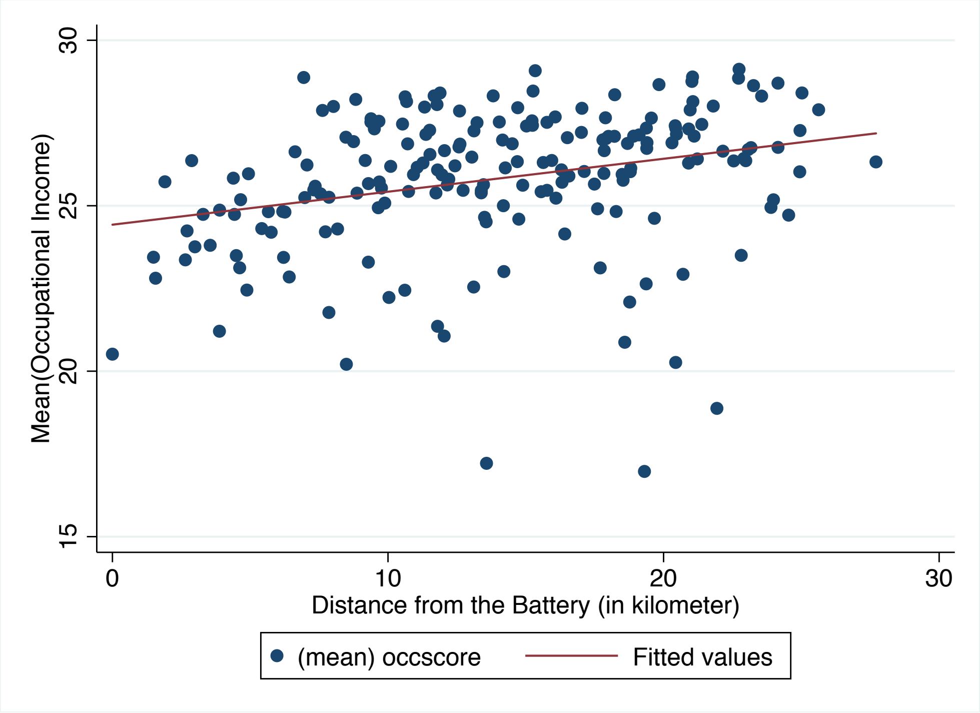


Figure 119: Mean Occupational Income

(f) 1930



(g) 1940

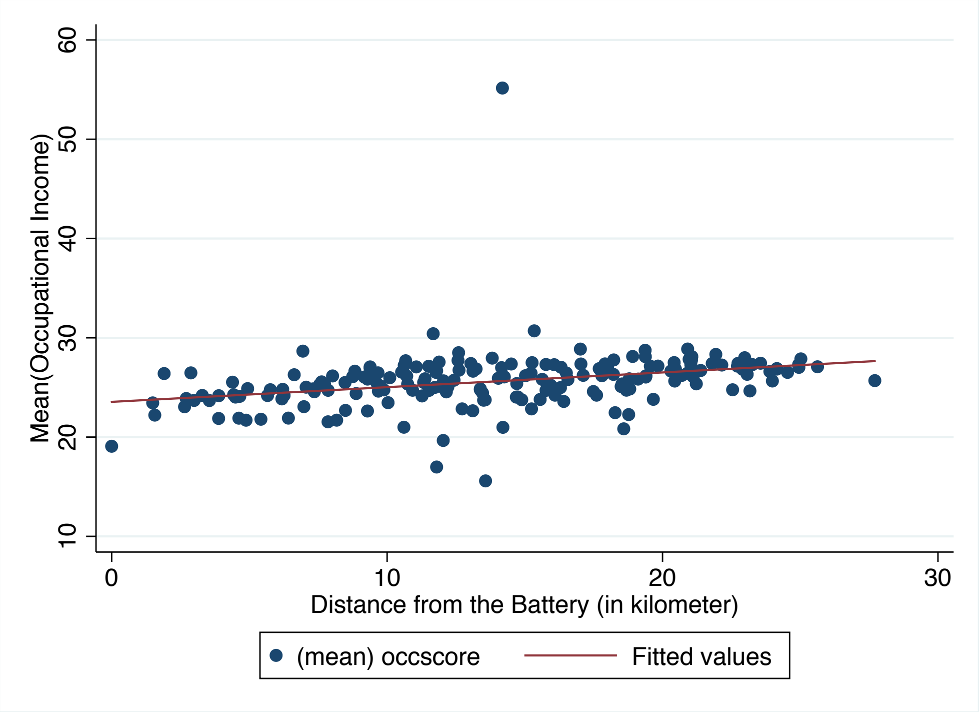
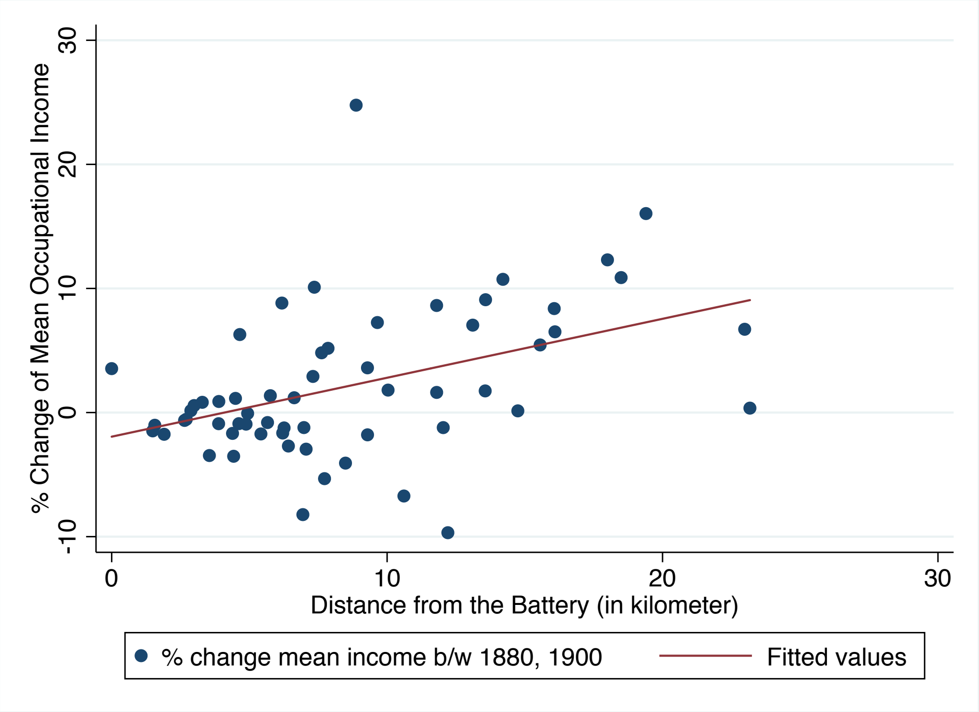


Figure 120: % Change of Mean Occupational Income

(a) 1880-1900



(b) 1900-1910

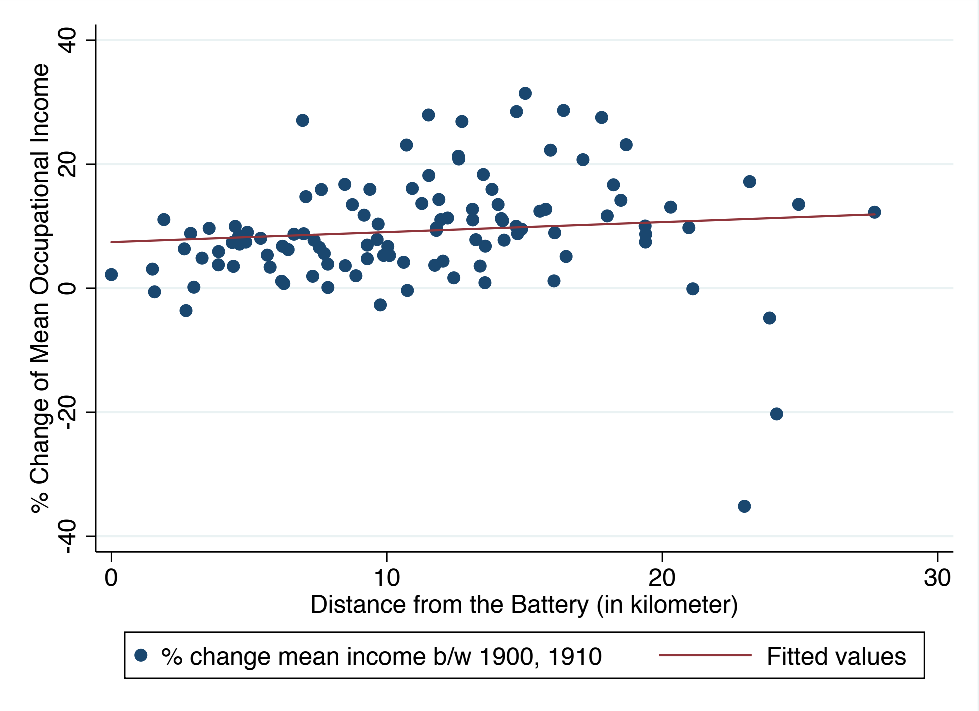
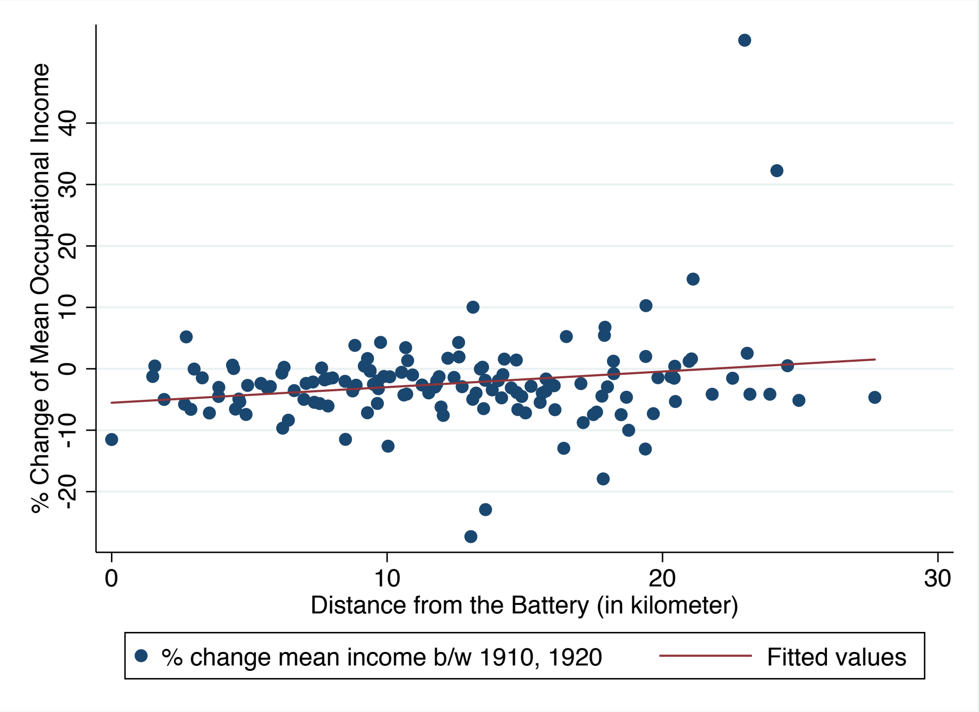


Figure 120: % Change of Mean Occupational Income

(c) 1910-1920



(d) 1920-1930

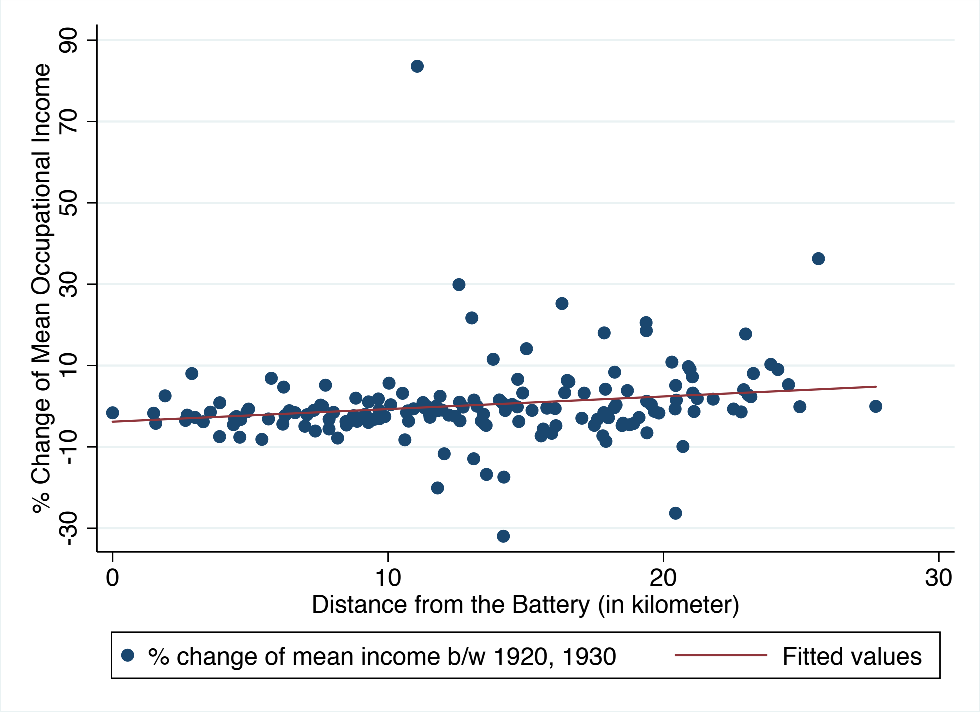
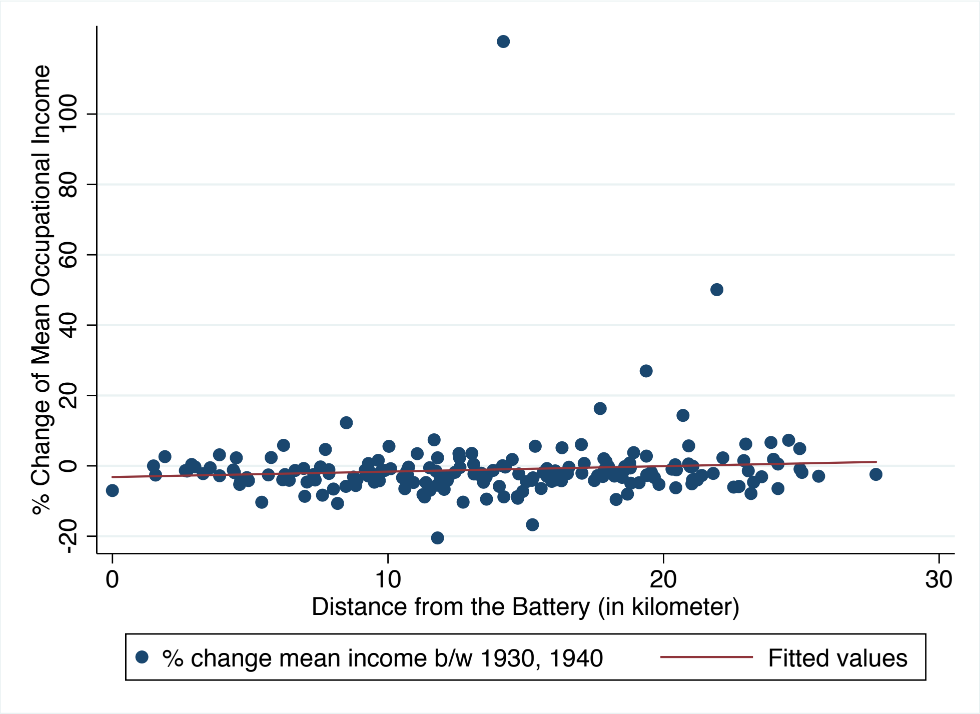


Figure 120: % Change of Mean Occupational Income

(e) 1930-1940



1.4.4.2 Decomposition of the various flows among the core, the periphery, and the rest of the world

Neighborhoods' income changes between two periods are composed of 6 factors: the relative difference between entrants and stayers; the change in income of stayers; the relative difference between leavers and stayers as well as the relative magnitudes of the flows. Section 1.4.2, 1.4.3 looks at the relative difference between leavers and stayers in the core and the relative difference between entrants and stayers in the periphery respectively in relation to Jackson (1985).

In this Section 1.4.4.2, I decompose various flows among the core, the periphery, and the rest of the country, along with associated incomes. As NTA-level results in Section 1.4.2, 1.4.3 show, flows within the metropolitan area are different from flows from outside the NYC-metropolitan area. For example, in terms of city core leavers, the original neighborhood residents leaving to the periphery of the city may be different (in terms of income and other characteristics such as age and race) from residents leaving to NYC metropolitan area such as Westchester County, or residents moving to outside NYC metro area entirely. Similarly, entrants to the periphery neighborhood in the city could be from another neighborhood in the city including the core of the city (as Jackson (1985) discussed), or from NYC metro area (e.g. Westchester county located in the north of NYC which is a part of NYC metro area), or from outside NYC metro area.

Specifically, I decompose the changes in the core and the periphery in the following way. In terms of the core, I look at various flows including the relative income difference between leavers and stayers as well as the corresponding relative magnitudes of those flows at the neighborhood level. In terms of the periphery, I look at various flows including the relative income difference between entrants and stayers as well as the corresponding relative magnitudes of those flows at the neighborhood level.

These decomposition analyses are complementary to results in Table 14, 16, 18, 110, 112 for leavers (relative to stayers at the neighborhood level) at the core of the city, and in Table 114, 116, 118, 120, 122 for entrants (relative to stayers at the neighborhood level) at the periphery of the city.

I look separately at flows within the city, flows within the metro area, and flows from the outside the metro area in analyzing neighborhood changes at the neighborhood level.

- Decomposition of the core of the city (leavers and stayers at the core of the city)

I decompose various flows of the core of the city including the relative income difference between leavers and stayers as well as the corresponding relative magnitudes of those flows at the neighborhood level. In Figure 121, the size of hollow circle denotes that relative ratio of such migration-type, whereas y-axis indicates the mean occupational income of each migration type. In Figure 121, as x-axis captures the number of flows, strictly positive sign indicates inflows (entrants) whereas strictly negative sign indicates outflows (leavers), and the distance away from 0 captures the total number of flows across migration-type.

Throughout the study period, neighborhood stayers at the core (in Lavender in Figures 121a, 121c, 121e, 121g) had the highest mean occupational income relative to other neighborhood leavers at a varying degree. Related to Jackson (1985), core leavers to the periphery in the city (in Cranberry) and to NYC metro area (in Orange) relative to the core stayers matter. Figures 121a, 121c, 121e, 121g show the core stayers' income (in Lavender) was higher than those two groups throughout the study period. This is consistent with my findings in Section 1.4.2 that it was not the rich who left the center of the city. However, although the neighborhood stayer at the core may have been the richest, only a small fraction of people stayed in the same neighborhood and the majority of them left the core of the city — some migrated to the periphery of the city, others migrated to NYC metro area, and the others migrated outside the NYC metro area at all.

The relative magnitude of flows gives us a richer story regarding the evolution of the core of the city. First of all, till 1910, the proportion of leavers who are leaving to outside NYC metro area (in Teal) was higher than any other group, whereas starting in 1920, the proportion of neighborhood leavers moving to the periphery in the city became higher than any other group. This implies the magnitude of within-city internal migration increased greatly around 1920 which was at the peak of intra-city transit infrastructure investment after the Dual Contract. It is also notable that between Years 1920 and 1930, the magnitude of leavers to the periphery of the city and is astonishing — the number of leavers who were leaving the core of the city were almost five times bigger than the number of entrants to the core — implying that population decline in the core of the city was more dramatic than ever.

Therefore, although Jackson (1985)'s fundamental claim about the growth of high income at the edge relative to the center holds true for my study (which I will discuss in the following decomposition), Jackson (1985)'s claim of the rich leaving the center of the city as the mechanism for explaining the growth of high income at the edge does *not* hold true for my study. Jackson's straightforward inference of the rich leaving the center and moving to the periphery does not apply to my longitudinal data-based migration analysis. To recap, the core stayers were richer than any other leaver groups at any destination, and it was not the rich who left the core of the city.

The flows of entrants also capture how the change — change of incomes at the center declining relative to the center — actually happened. The core entrants' income was much lower than leavers and this is especially pronounced for core entrants from outside NYC metro area (in Maroon). Entrants from outside NYC metro area were the largest entrant group with the lowest mean income, and therefore, this must have caused the core of the city's income to decrease.

To recap, the transit changes in my study period had a similar nature of Jackson (1985) and the postwar period and incomes at the edge were rising relative to the center. However, the mechanism behind these changes was *not* a simple shuffling of the rich and poor.

- Decomposition of the periphery of the city (entrants and stayers at the periphery of the city)

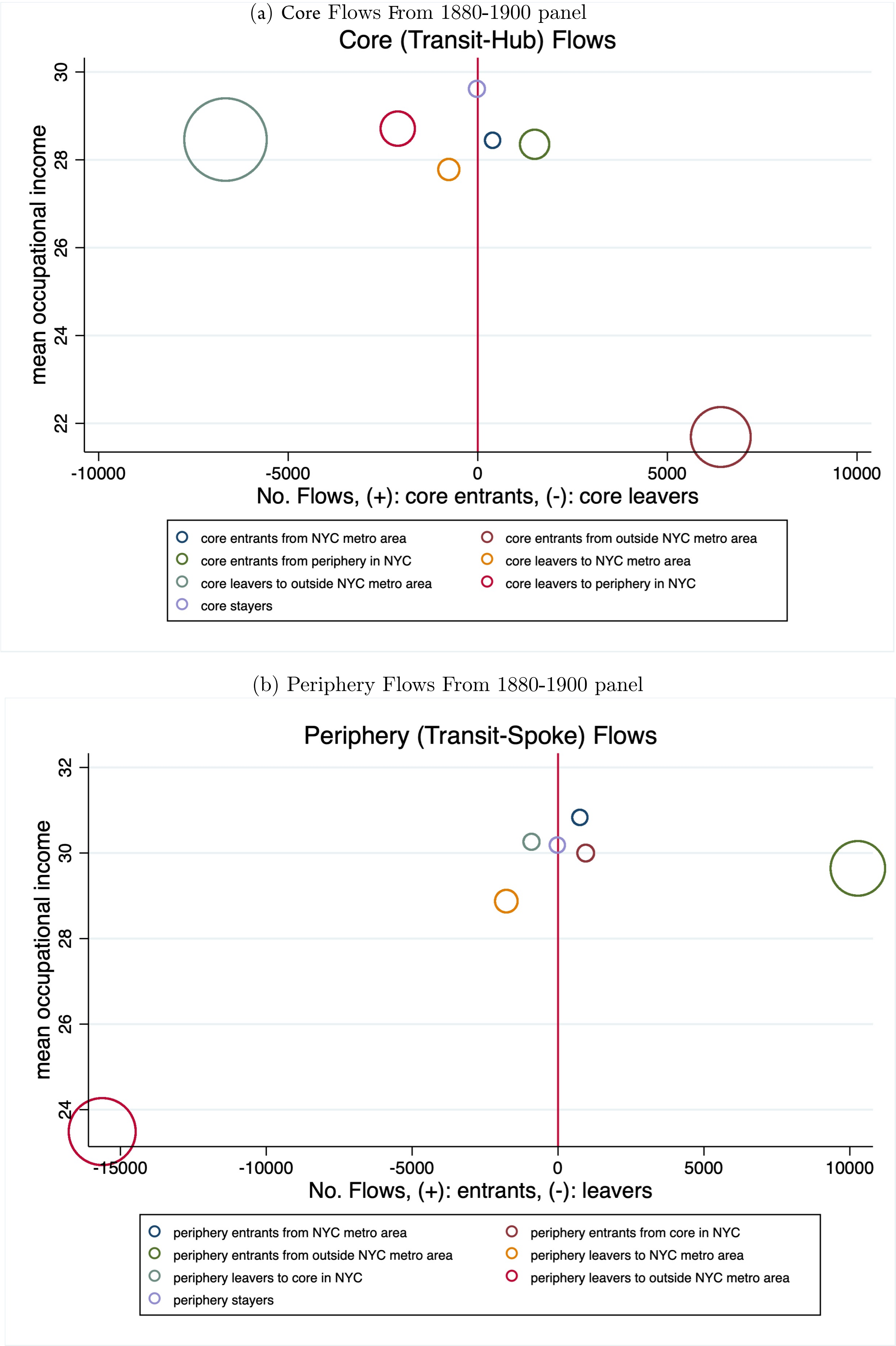
I also decompose various flows of the periphery of the city including the relative income difference between entrants and stayers as well as the corresponding relative magnitudes of those flows at the neighborhood level. Throughout the study period, neighborhood stayers at the periphery of the city (in Lavender in Figures 121b, 121d, 121f, 121h, 121j) had a higher mean occupational income than other periphery entrant groups. Related to Jackson (1985), the income of periphery entrants from the core (in Maroon) relative to periphery stayers matter. Figures 121b, 121d, 121f, 121h, 121j show that income of the periphery entrants from the core (in Maroon) was lower than that of periphery stayers (in Lavender) throughout my study period.

Therefore, Jackson (1985)'s straightforward inference about the rich from the core moving to the periphery as the primary mechanism for the edge's relative income growth does not hold for my study. The periphery entrants from the city core had lower mean income than periphery stayers, and their relative magnitude in terms of the number of people was fairly small. The primary mechanism behind the income growth are three forces: 1. the periphery leavers moving to outside New York metro area had lowest mean income than any other group, and they left the city periphery greatly (in Cranberry), 2. periphery entrants from outside NYC metro area (in Green) had much higher mean income than periphery leavers and magnitude of this inflow was big enough to dominate all the other forces, 3. the periphery stayers' income increased substantially (Figure 122b).

To summarize, incomes at the periphery were rising relative to the center due to entrants from outside NYC metro area with a higher income than periphery leavers and this flow was sizable both in terms of the relative income difference between leavers and entrants as well as the relative magnitudes of the flows.

Finally, the decomposition analyses show that the neighborhood stayers at the periphery were not poorer than most entrants into the periphery. However, Table 114, 116, 118, 120, 122 show that the new suburbanites at the periphery were not richer than the people who stayed at the periphery.

Figure 121: Mean Income and Magnitude of Flows Across Migration-type: 1880-1900

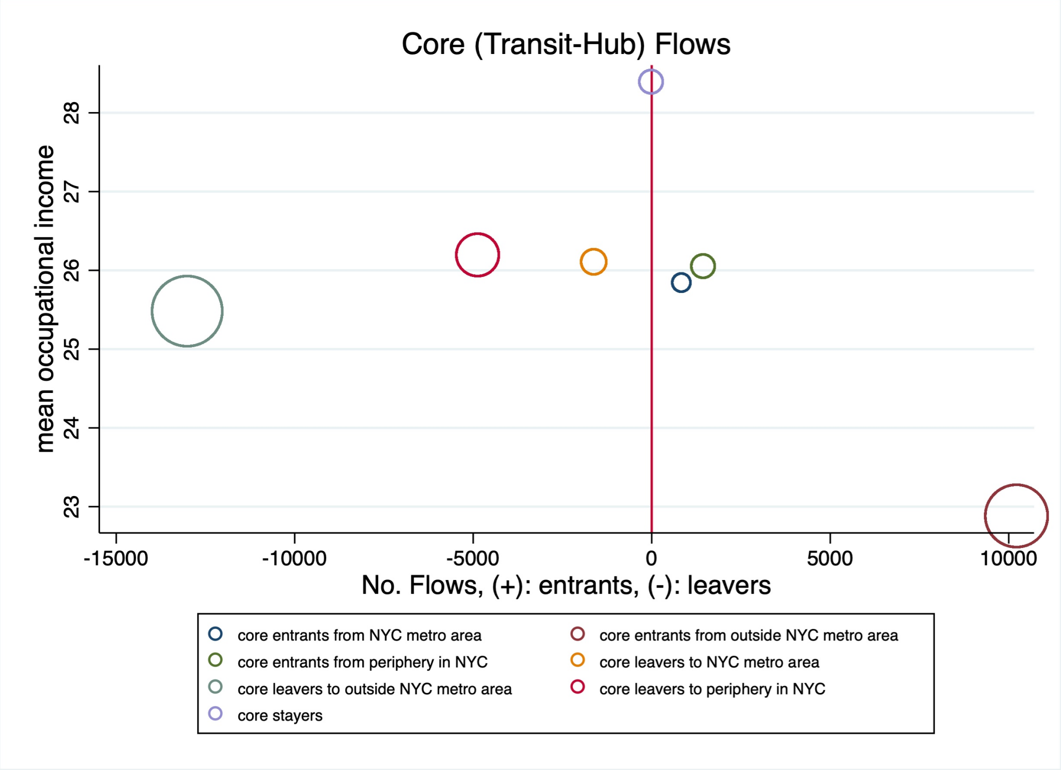


(a) Core Flows From 1880-1900 panel

(b) Periphery Flows From 1880-1900 panel

Figure 121: Mean Income and Magnitude of Flows Across Migration-type: 1900-1910

(c) Core Flows from 1900-1910 panel



(d) Periphery Flows From 1900-1910 panel

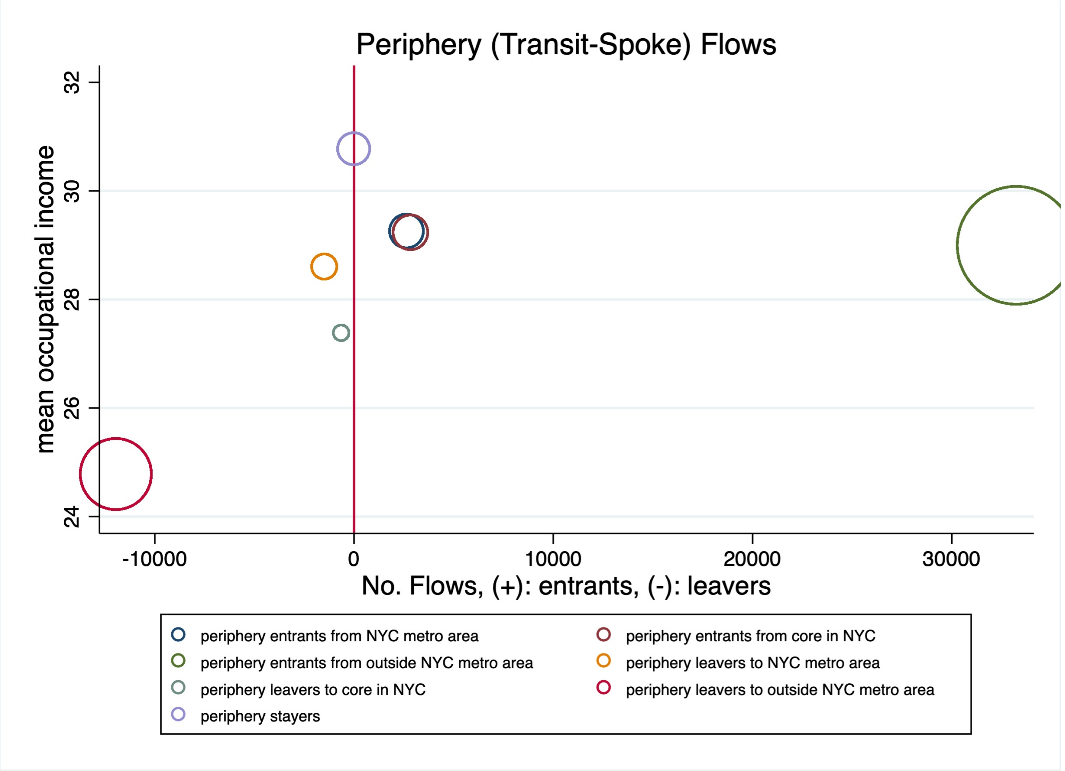
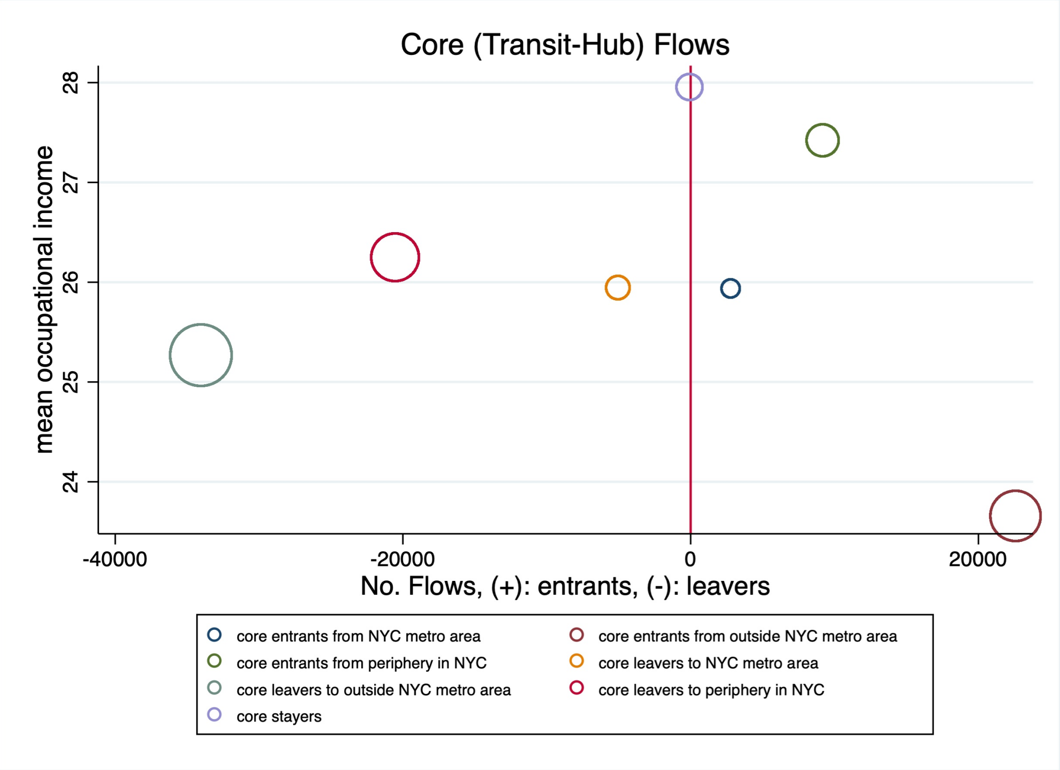


Figure 121: Mean Income and Magnitude of Flows Across Migration-type: 1910-1920

(e) Core Flows From 1910-1920 panel



(f) Periphery Flows From 1910-1920 panel

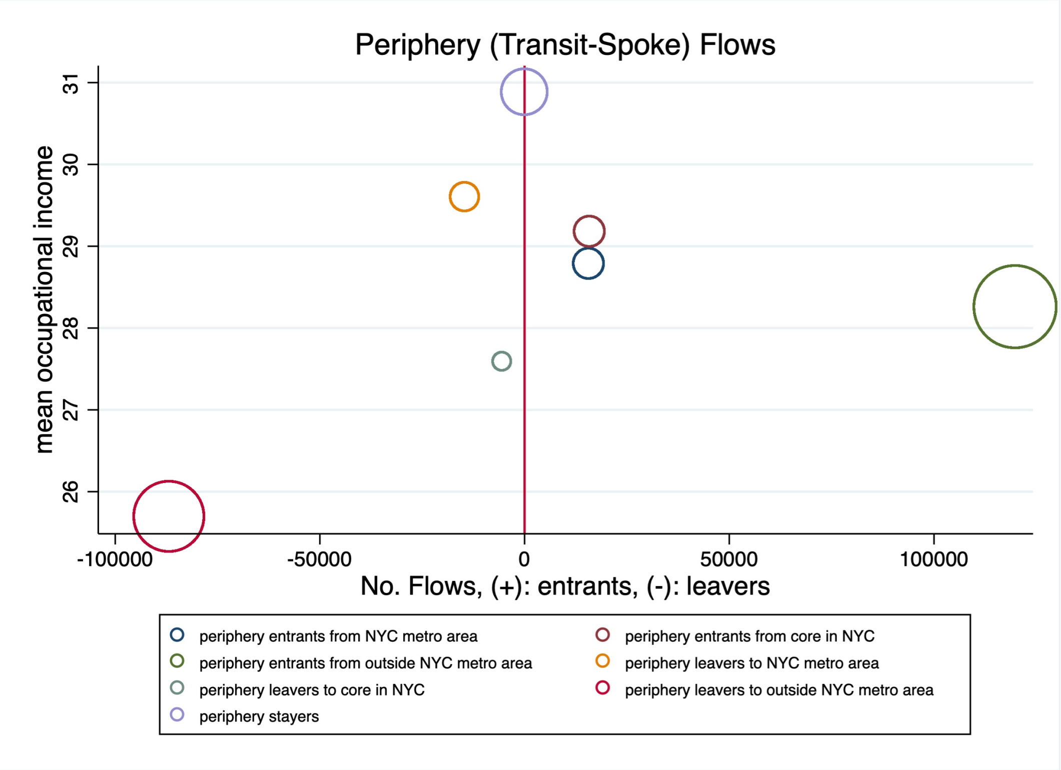
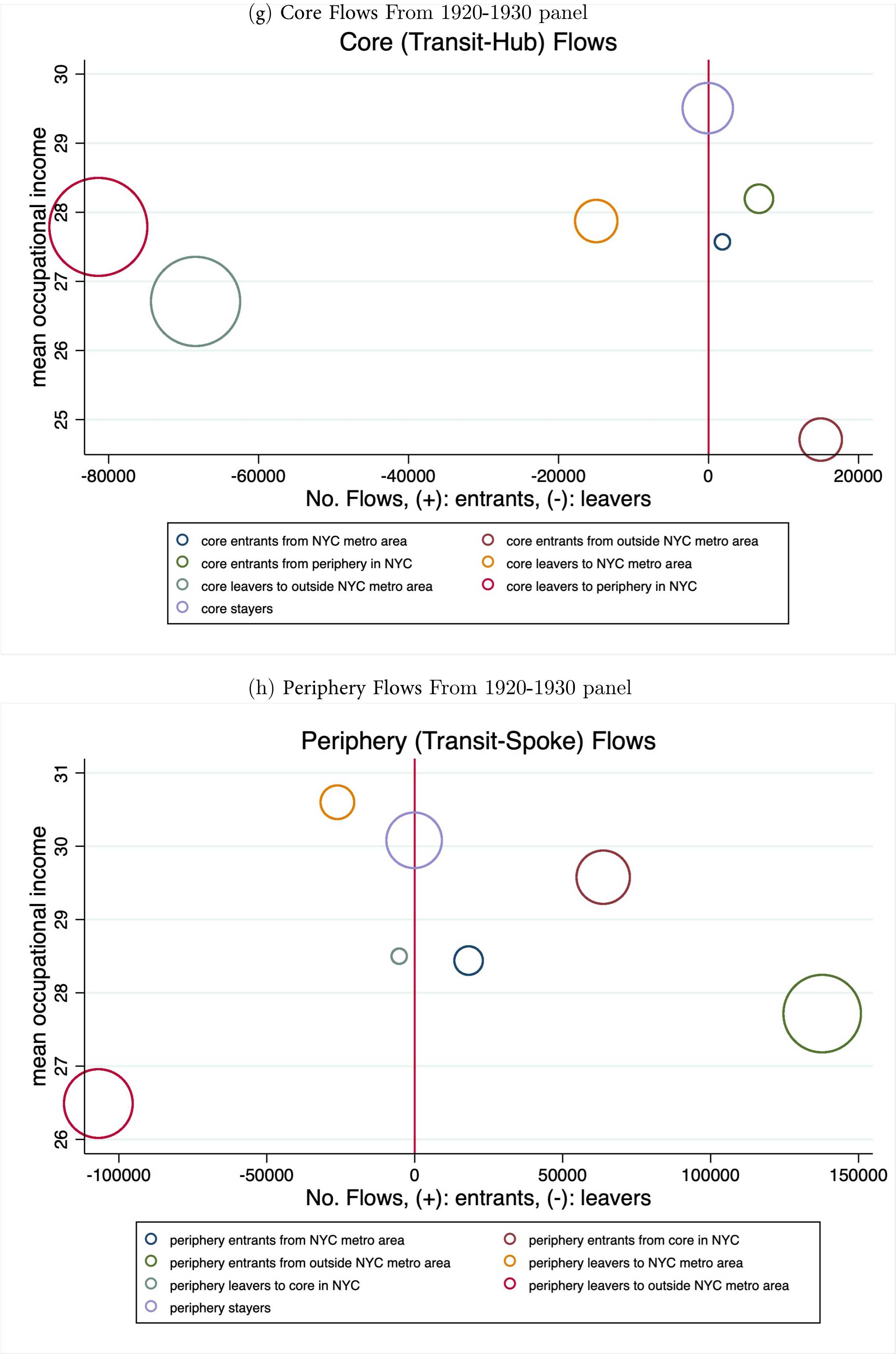


Figure 121: Mean Income and Magnitude of Flows Across Migration-type: 1920-1930

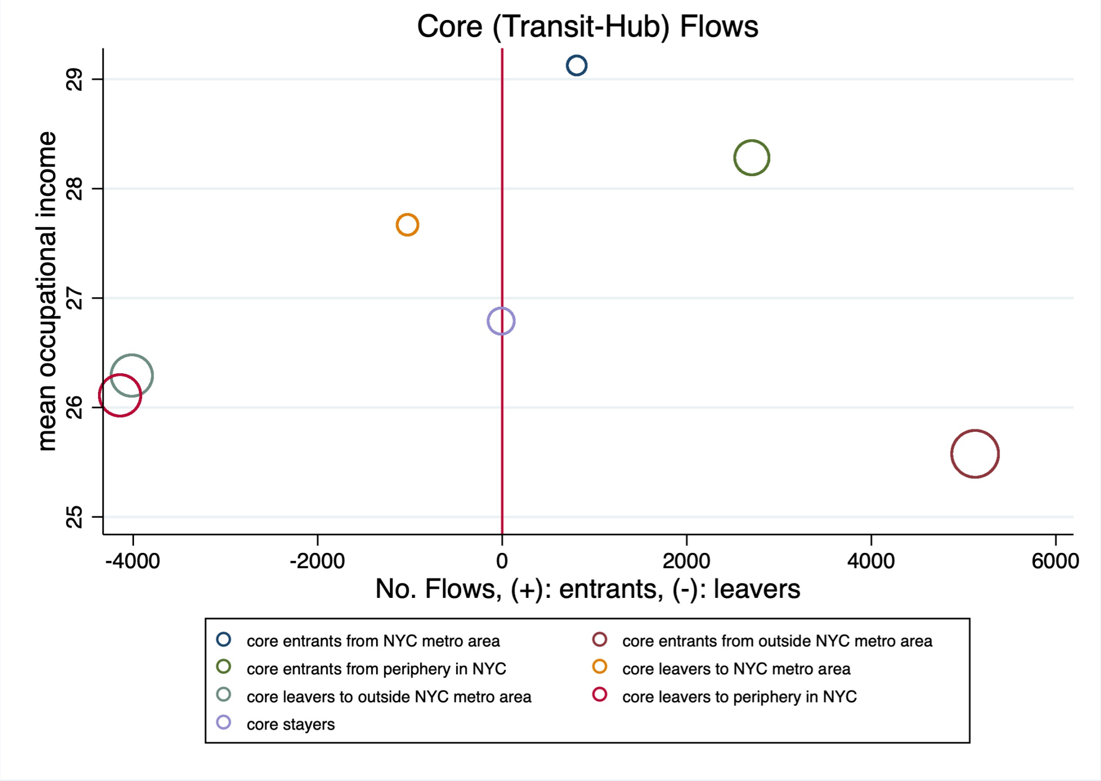


(g) Core Flows From 1920-1930 panel

(h) Periphery Flows From 1920-1930 panel

Figure 121: Mean Income and Magnitude of Flows Across Migration-type: 1930-1940

(i) Core Flows From 1930-1940 panel



(j) Periphery Flows From 1930-1940 panel

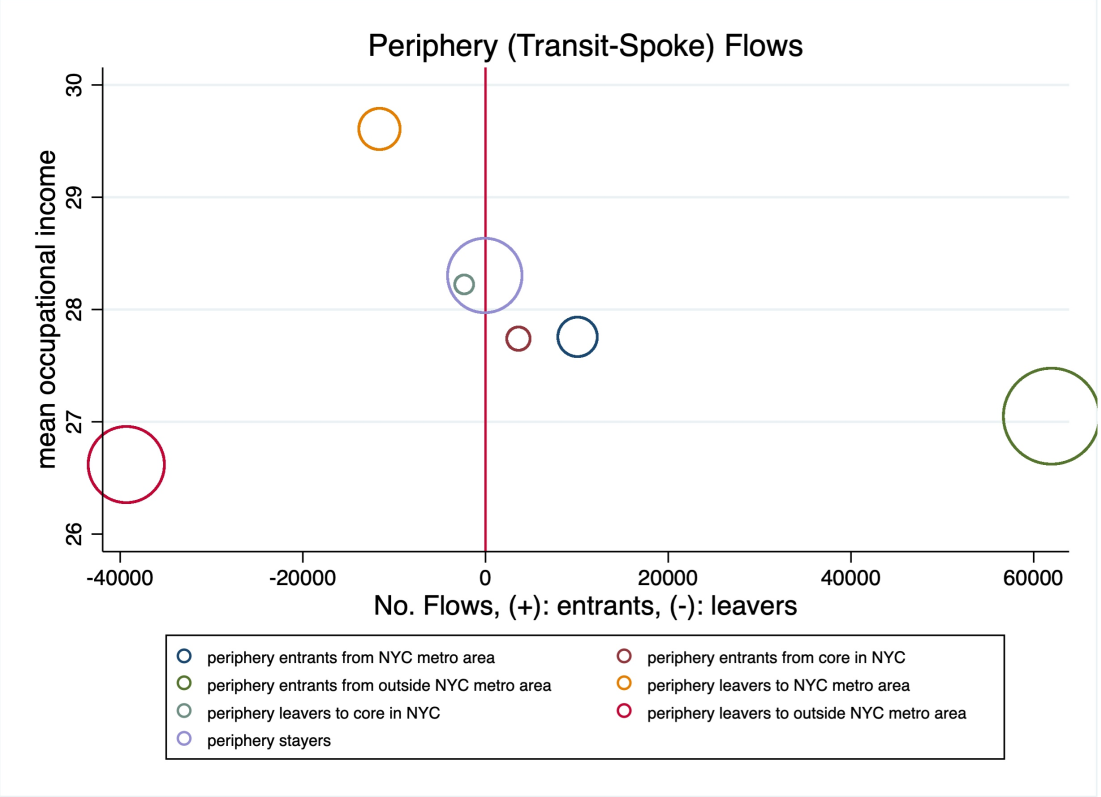
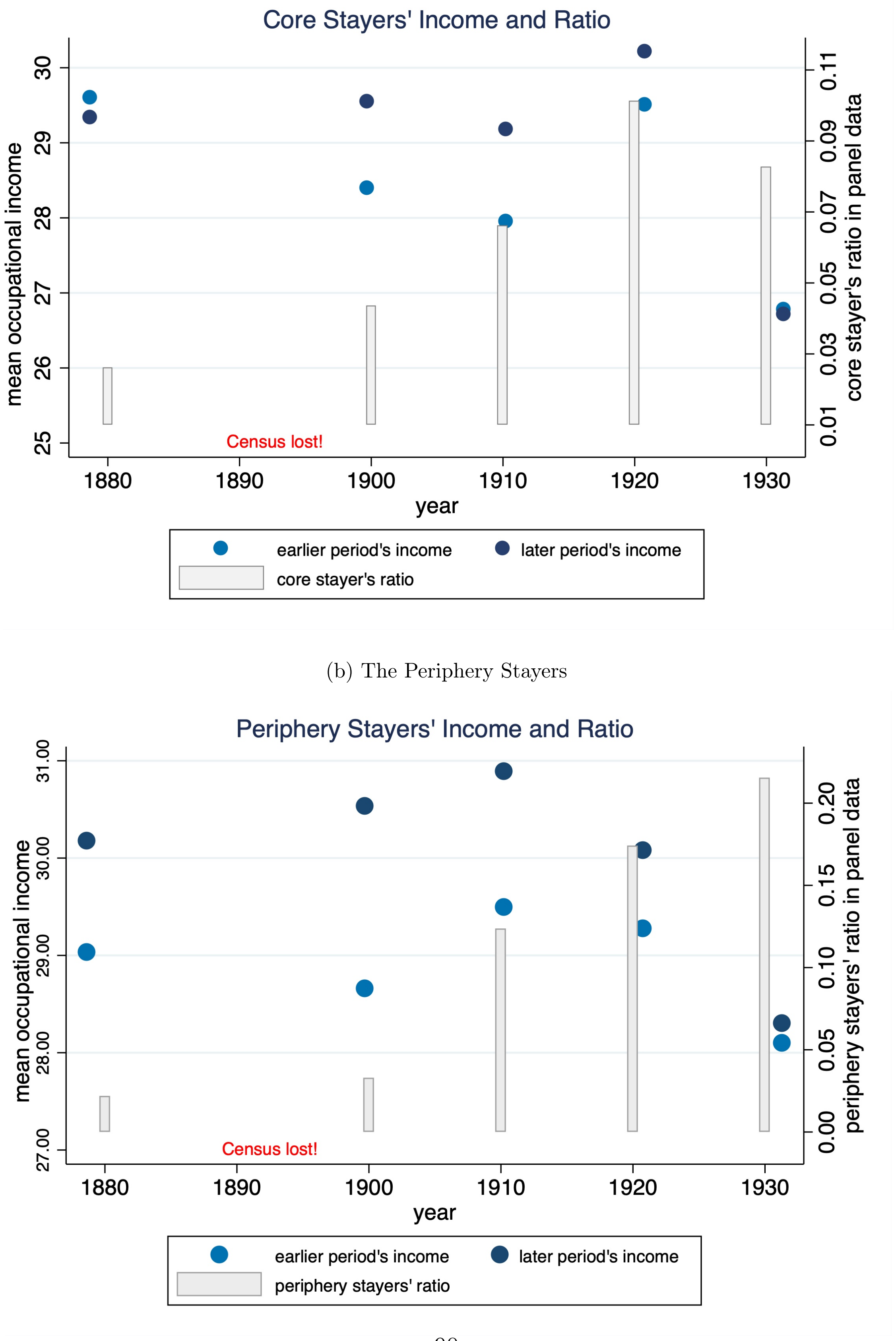


Figure 122: The Core and Periphery Stayers' Income and Ratio



(a) The Core Stayers

(b) The Periphery Stayers

1.5 Conclusion

With the 21st-century advanced techniques and computational power, I construct longitudinal database by linking complete-count US census records from 1870 to 1940. I analyze income and socioeconomic status of individuals who lived in the New York City and analyze whether the dynamic process of suburbanization in New York systematically differed for the poor and rich (and other characteristics).

Longitudinal data reveal that in the core, it was not the case either that rich people left or that poor people stayed; in the periphery, people who moved into the periphery were not richer than original residents. The suburbanization in prewar New York was probably different from postwar suburbanization. The people who lived in the central city were not poor; those who left were not more affluent.

My study period also captures the growth of high income at the periphery relative to the center as in Jackson (1985)'s period (1815-1875) and the postward period. However, unlike other works that do not use panel data, I use panel data of individuals to uncover the mechanism behind the growth of high income at the edge relative to the center. Instead of making an inference about who moved, left, and stayed, I show how the change actually happened.

Essentially, the transit infrastructure improvements and changes in my study period had the same nature of Jackson's period and the post war period—incomes at the edge rising relative to the center. However, the anatomy of how this change actually happened shows that incomes rising at the edge (relative to the center) was not a simple shuffling of rich and poor. Up until the Great Depression, flows of migrants from and to outside the NYC metropolitan area were the dominant force in changing average income. Richer people from outside migrated to the periphery, whereas poorer people from outside migrated to the core. The people from the core who left the NYC metropolitan area entirely were richer than the people from the periphery who left the NYC metropolitan area. Finally, people who stayed at the periphery got richer as the metropolis grew.

Finally, while this paper quantitatively shows the dynamic process of suburbanization in relations to transit infrastructure improvement, it still is not clear whether the welfare consequences differ across the rich and the poor (or high skill and low skill workers). Analyzing rigorous welfare comparisons offers a direction for future research.

Chapter 2

European Immigrants and the United States' Rise to the Technological Frontier

Costas Arkolakis, Michael Peters, Sun Kyoung Lee[1]

[1] We thank Donald Davis, Dan O'Flaherty, Sam Kortum, Ellen McGrattan, Ferdinando Monte for excellent suggestions and Xiaoyun Qiu and Jack Liang for outstanding research assistance. The authors would like to thank NSF for support under RIDIR grant #1831524 as well as NBER small enterprenurship grant. All errors are our own.

European Immigrants and the United States' Rise to the Technological Frontier

Costas Arkolakis
Yale

Sun Kyoung Lee
Yale

Michael Peters
Yale

Abstract

What is the role of immigrants on (American) Growth? To answer this perplex question, we undertake a massive effort of collecting, digitizing, and harmonizing micro and macro economic data from the 19th and early 20th century. The data originate from the historical manufacturing and demographic census of the United States, immigration records datasets and the universe of US patents. To analyze the counterfactual implications of alternative allocations of immigrants, we develop a dynamical trade model where heterogenous firms make innovation and exporting decisions across space and time. The model predicts that the timing and the spatial allocation of immigrant arrivals affect the path of growth outcomes for each location and the aggregate US economy. We use the structural equations arising from the model to interpret empirical findings from difference-in-difference analysis for the importance of the influx of skilled immigrants on the differential growth of US counties. Counterfactual scenarios of alternative allocation of skilled immigrants from different countries across space and time reveal the economic impact of barriers to migration to the United States economy.

2.1 Introduction

The transformation of the US economy in the last two hundred years has been remarkable. While being primarily rural in the beginning of the 19th century, the US had developed into an essentially industrial nation when the century came to an end. More strikingly, after lagging behind the technological frontier (represented by the UK) for most of the 19th century, the US entered the twentieth century as the global technological leader and the richest nation on the globe (Gordon, 2017). During this time period, which is also referred to as the "The Second Industrial Revolution", the US economy also experienced a massive in-flow of immigrants, mostly from the European continent.

To answer this question, we develop and harmonize a massive dataset that combines longitudinal information on immigrants and their occupations along with measures of economic outcomes such as, output, wages, and innovation at the disaggregated county-industry level. Using this information we plan to investigate to what extent this influx of immigrants was an important contributor to the transformation of the American economic landscape between 1850 and 1940 and provide a comprehensive account of the importance of spatial mobility for the immigrants' life-cycle. To do we develop a dynamical model of innovation where firms in each location innovate taking into account the full stream of future incomes that can be generated there now and in the future. Firms hire production and research workers. Workers, immigrants or natives, can specialize in either of these activities depending on their idiosyncratic abilities that are determined by their comparative advantage. The arrival of immigrants leads to standard scale effects, as in the Romer (1990) model or as we formally argue may even futher boost innovation if immigrants have a comparative advantage on this activity. The model allows the analysis of a range of regional policies, such as one ones that improve labor productivity, improve innovation productivity improvements, or increase population due to arrivals of natives or immigrants.

We use original immigration records and historical passenger lists from the ships heading from Europe to the US.[1] These data sources are real treasure troves for empirical researchers, as they contain direct micro data on immigrants' *pre*-migration occupations.[2] In particular, both in the immigration records and in the passenger lists, all immigrants were required to give a detailed account of their last occupation in Europe along with other important information such as the time of arrival in the US, the place of residence in Europe, and age. We link these immigration datasets with the restricted-use complete count US Federal demographic decennial censuses from 1850-1940 using modern record-linking techniques. To link these datasets, we exploit the fact that both the immigration records (and passenger lists) and the federal US demographic Census contain time-invariant individual information. In that way we construct a large-scale micro *panel* data set for immigrants with information on their *pre*-migration occupations and their *post*-migration labor market outcomes and spatial mobility patterns over their whole life-time.

Using this integrated dataset, we study one specific mechanism by which immigration could have led to American prosperity: the transmission of *new* knowledge. While anecdotal evidence of this channel, whereby European immigrants brought novel ideas to the American shore, is abundant, there is no systematic evidence whether this mechanism was quantitatively important. To fill that gap, we first show that the information on immigrants' pre-immigration occupations allows us to construct empirical proxies for the flow of novel ideas, or in short, *flows of knowledge*. We then combine this immigrant micro data with novel measures of wages, productivity growth and patent activity. In particular, by digitizing the published results for the historical Manufacturing Census from 1860 to 1939, we construct data on wages and productivity measures at the county-industry level. Secondly,

[1] The immigration database of 13 million immigrants and the passenger lists of around 5 million immigrants leaving for the US via the German port Hamburg, the so-called "Hamburg Passenger Lists" were provided to us for research purposes by the Battery Conservancy and the Archives of the city of Hamburg, respectively. See http://www.castlegarden.org and http://www.germanroots.com/hamburg.html for additional information. To the best of our knowledge, these data sources have yet to be used in empirical research.

[2] We constructed a crosswalk between these published occupational strings and the *Historical International Classification of Occupations (HISCO)*. For more details on the Historical International Classification of Occupations, see http://historyofwork.iisg.nl/index.php.

we also complement our data with new county-industry measures of patent activity and patent novelty.

This combination of micro-information on immigrants and macro-measures of productivity and spatial idea creation allows us to relate knowledge flows (as proxied by inflowing immigrants with pre-migration expertise) to data on productivity growth and patent activity for the study period. By doing so, we provide novel evidence on potential mechanisms by which past immigrant settlements could affect economic outcomes (see e.g. Nunn et al. (2017) and Akcigit et al. (2017)). Furthermore, our study period is not only interesting in itself, but it also provides an ideal laboratory to empirically identify the importance of idea flows, which feature prominently in recent theories of economic growth (Kortum, 1997; Lucas Jr and Moll, 2014; Perla and Tonetti, 2014). As communication flows and technology were far less developed than those of today, the importance of embodied knowledge transmission was arguably much more important at that time.

Furthermore, we investigate the role of spatial mobility for the immigrants' earnings life-cycle. We exploit the longitudinal and spatial aspects of our dataset more intensely. In particular, we construct "spatial-sector"-based earning measures from our newly digitized information on manufacturing wages. In fact, it might be at the heart of understanding earnings differences between natives and immigrants (see eg Abramitzky et al. (2012a)) as our data shows that there a is systematic positive correlation between average wages, urbanization and immigrant shares in the 19th century.[3]

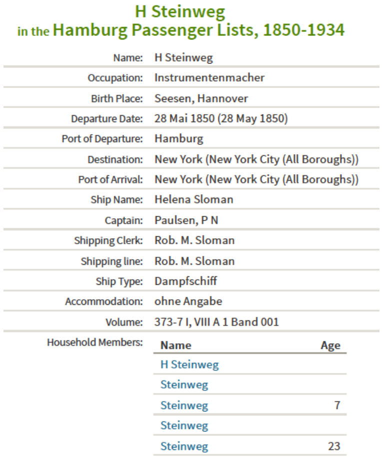
The rest of the paper is structured as follows. Section 2.8 2.8 concludes.

[3] This finding is consistent with the findings of a literature in urban economics that finds large city-wage premia in recent data (see e.g. Roca and Puga (2017)).

2.2 A guiding example: The Story of Heinrich Engelhard Steinweg

To illustrate how our novel data can be merged with existing data sources to shed light on the process of technology transfer from old Europe, we start by a particular example: the case of Heinrich Engelhard Steinweg, later known as Henry Engelhard Steinway, the founder of renowned piano manufacturing company *Steinway & Sons*.

Heinrich Steinweg left Germany on May 28th, 1850 via the port of Hamburg. This information is declared on the *Hamburg Passenger Lists (1850-1934),* which is available to us through a cooperation with the Archives of the city of Hamburg. His shipment record also indicates his pre-immigration occupation in Germany as *Instrumentenmacher* (instrument maker).[4] As we can see from the same record, shown in Figure 21, his destination was New York and he was accompanied by four family members.



H Steinweg
in the Hamburg Passenger Lists, 1850-1934

Name:	H Steinweg
Occupation:	Instrumentenmacher
Birth Place:	Seesen, Hannover
Departure Date:	28 Mai 1850 (28 May 1850)
Port of Departure:	Hamburg
Destination:	New York (New York City (All Boroughs))
Port of Arrival:	New York (New York City (All Boroughs))
Ship Name:	Helena Sloman
Captain:	Paulsen, P N
Shipping Clerk:	Rob. M. Sloman
Shipping line:	Rob. M. Sloman
Ship Type:	Dampfschiff
Accommodation:	ohne Angabe
Volume:	373-7 I, VIII A 1 Band 001

Household Members:

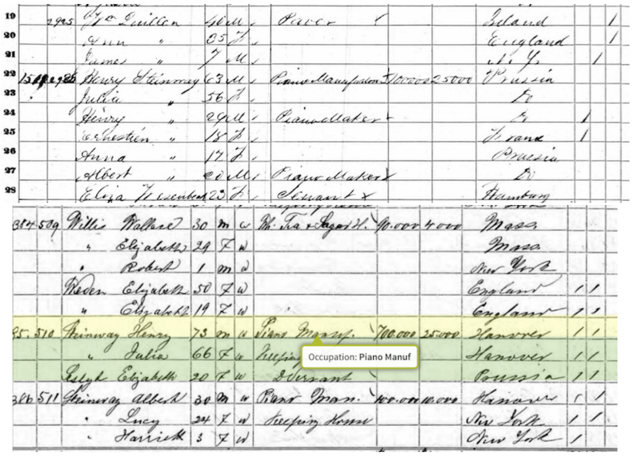
Name	Age
H Steinweg	
Steinweg	
Steinweg	7
Steinweg	
Steinweg	23

Figure 21: Heinrich Steinweg in the *Hamburg Passenger Lists, 1850-1934*

We can track Mr. Steinweg, now Mr. Steinway, in the subsequent US population cen-

[4] Henry Steinway started working on producing pianos early on with immediate success. But the unstable political climate following the revolutions of 1848 and the limited economic opportunities for a man working outside a guild let him to immigrate to the US. See Claudius Torp, "Heinrich Engelhard Steinway." In Immigrant Entrepreneurship: German-American Business Biographies, 1720 to the Present, vol. 2, edited by William J. Hausman. German Historical Institute.

suses, shown in Figure 22. Both in in 1860 and 1870, Mr. Steinway and his family are recorded to reside in New York. Furthermore, his occupation *piano manufacturer* indicates Steinway's successful transition from a piano maker in Germany to the piano manufacturer of the US.



19		2925	Jno Quillen	40	M		Paver			Ireland		1
20			Ann "	35	F					England		1
21			James "	7	M					N.Y.	1	
22	1500	2926	Henry Steinway	63	M		Piano Manufacturer	100000	25000	Prussia		
23			Julia "	56	F					Do		
24			Henry "	29	M		Piano Maker			Do	1	
25			Ernestine "	18	F					France	1	
26			Anna "	17	F					Prussia		
27			Albert "	20	M		Piano Maker			Do		
28			Eliza Wiesenbeck	23	F		Servant			Hamburg		

384	509	Willis Wallace	30	M	W	Wh. Tea & Sugar H.	10.000	4000	Mass		
		" Elizabeth	29	F	W				Mass		
		" Robert	1	M	W				New York		
		Weden Elizabeth	50	F	W				England	1	1
		" Elizabeth	19	F	W				England	1	1
385	510	Steinway Henry	73	M	W	Piano Manuf.	700.000	25.000	Hanover	1	1
		" Julia	66	F	W	Keeping			Hanover	1	1
		Lelyh Elizabeth	20	F	W	D Servant			Prussia	1	1
386	511	Steinway Albert	30	M	W	Piano Man.	100.000	10.000	Hanover	1	1
		" Lucy	24	F	W	Keeping House			New York	1	1
		" Harriett	3	F	W				New York	1	

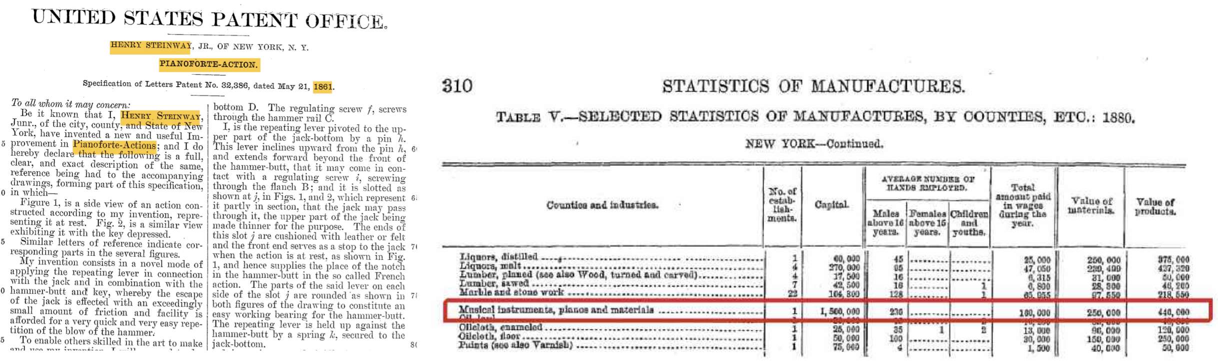
Notes: The figure shows Mr Steinway's US census records in 1860 (left panel) and 1870 (right panel).

Figure 22: Henry Steinway in the *US Census Schedules 1860 and 1870*

That this successful career trajectory might have been in part due to Mr Steinweg's prior knowledge is consistent with micro data on patenting. Using digitized historical patent data from the United States Patent and Trademark Office, we could extract a number of patents granted to him and his sons' names. For example, Steinway's famous piano-forte patent, dated 1862, is shown in the left panel of Figure 23.

Finally, we can use our newly digitized data from the US Census of Manufactures to learn about the economic magnitude of Mr. Steinway's success. While the US Census of Manufacturers data is not at the plant level (but reported at industry-by-county cells), the information is detailed enough to identify the main manufacturing plant of Steinway & Sons in Queens, NY. As the right panel of Figure 23 shows, the Steinway family had an enormous impact on manufacturing production in the New York area. The digitized Census of Manufacturers for the year 1880, for example, reveals that this single piano manufacturing plant was one of the most capital intensive sectors in New York City with more than $1.5

millions of capital and sales close to a half a million dollars.



UNITED STATES PATENT OFFICE.

HENRY STEINWAY, JR., OF NEW YORK, N. Y.

PIANOFORTE-ACTION.

Specification of Letters Patent No. 32,386, dated May 21, 1861.

To all whom it may concern:

Be it known that I, Henry Steinway, Junr., of the city, county, and State of New York, have invented a new and useful Improvement in Pianoforte-Actions; and I do hereby declare that the following is a full, clear, and exact description of the same, reference being had to the accompanying drawings, forming part of this specification, in which—

Figure 1, is a side view of an action constructed according to my invention, representing it at rest. Fig. 2, is a similar view exhibiting it with the key depressed.

Similar letters of reference indicate corresponding parts in the several figures.

My invention consists in a novel mode of applying the repeating lever in connection with the jack and in combination with the hammer-butt and key, whereby the escape of the jack is effected with an exceedingly small amount of friction and facility is afforded for a very quick and very easy repetition of the blow of the hammer.

To enable others skilled in the art to make

bottom D. The regulating screw *f*, screws through the hammer rail C.

I, is the repeating lever pivoted to the upper part of the jack-bottom by a pin *h*. This lever inclines upward from the pin *h*, and extends forward beyond the front of the hammer-butt, that it may come in contact with a regulating screw *i*, screwing through the flanch B; and it is slotted as shown at *j*, in Figs. 1, and 2, which represent it partly in section, that the jack may pass through it, the upper part of the jack being made thinner for the purpose. The ends of this slot *j* are cushioned with leather or felt and the front end serves as a stop to the jack when the action is at rest, as shown in Fig. 1, and hence supplies the place of the notch in the hammer-butt in the so called French action. The parts of the said lever on each side of the slot *j* are rounded as shown in both figures of the drawing to constitute an easy working bearing for the hammer-butt. The repeating lever is held up against the hammer-butt by a spring *k*, secured to the jack-bottom.

310 STATISTICS OF MANUFACTURES.

Table V.—SELECTED STATISTICS OF MANUFACTURES, BY COUNTIES, ETC.: 1880.

NEW YORK—Continued.

Counties and industries.	No. of establishments.	Capital.	Average number of hands employed.			Total amount paid in wages during the year.	Value of materials.	Value of products.
			Males above 16 years.	Females above 15 years.	Children and youths.			
Liquors, distilled	1	60,000	45	....	....	25,000	250,000	375,000
Liquors, malt	4	270,000	65	....	....	47,050	230,400	437,320
Lumber, planed (see also Wood, turned and carved)	4	17,500	16	....	....	6,315	31,000	50,000
Lumber, sawed	7	42,500	16	....	1	6,800	28,300	46,200
Marble and stone work	22	104,300	128	....	1	65,955	97,550	218,550
Musical instruments, pianos and materials	1	1,500,000	230	....	....	160,000	250,000	440,000
Oil, lead	[illegible]	[illegible]	[illegible]	....	2	[illegible]	[illegible]	[illegible]
Oilcloth, enameled	1	25,000	35	1	2	13,000	96,000	120,000
Oilcloth, floor	1	50,000	100	....	....	30,000	150,000	250,000
Paints (see also Varnish)	1	75,000	4	....	....	1,500	40,000	50,000

Figure 23: Mr Steinway's Pianoforte patent and Steinway & Sons in the *US Census of Manufacturers 1880*

In this proposal, we plan to analyze to what extent these patterns could be just one very special example of an extraordinary immigrant or whether there exist more systematic forces at play.

2.3 Primary Data Sources and Methodology

In this section we give more details on the data collection. In Section 2.3.1 we describe our main historical data sets, in Section 2.3.2 we discuss how we harmonize and link these data sources. An overview is contained in Figure 24. We rely on four distinct datasets. First of all, we digitize and harmonize two novel historical datasets: historical immigration records to construct a new database of millions of immigrants who arrived in the US during the Era of Great Migration and historical Census of Manufacturers data at the county/city-industry level to measure output and productivity.

We then combine these datasets with two other large datasets. Using record matching techniques, we link the data on immigrants with individual-level data from the US Population Census. This step yields an unprecedented panel dataset with pre- and post-immigration labor market outcomes for millions of immigrants. To construct direct measures of innova-

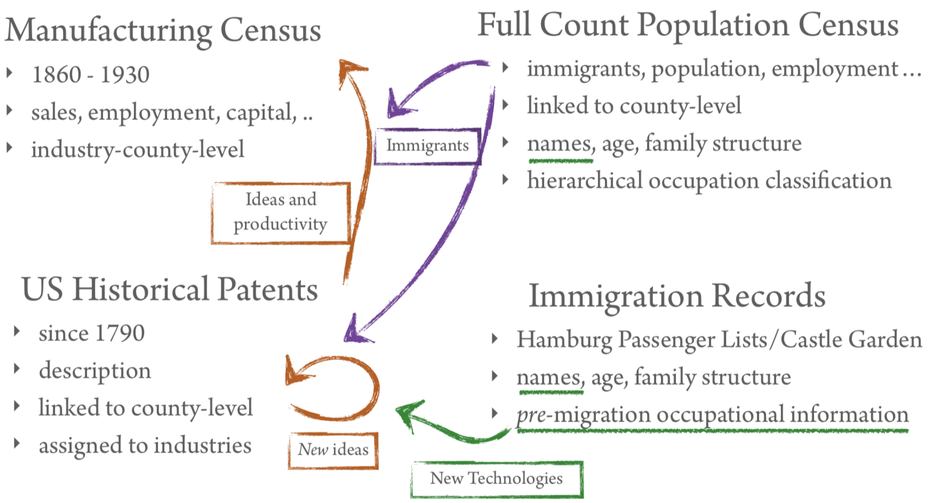

Manufacturing Census
1860 - 1930
sales, employment, capital, ..
industry-county-level
Full Count Population Census
immigrants, population, employment…
linked to county-level
names, age, family structure
hierarchical occupation classification
Immigrants
Ideas and productivity
US Historical Patents
since 1790
description
linked to county-level
assigned to industries
Immigration Records
Hamburg Passenger Lists/Castle Garden
names, age, family structure
pre-migration occupational information
New ideas
New Technologies

Figure 24: Data Construction: Combining the Data Sets

tive activity, we also incorporate the population of US historical patents since 1790. The combination of these datasets allows us to systematically explore the relationship between immigrants' prior expertise, productivity growth and patent activity at the county-industry level.

2.3.1 Historical Data on Immigrants, Local Productivity and Innovation

In this section we describe the four datasets, which form the basis of our analysis in detail.

2.3.1.1 *The Immigration Database: Immigration Records and Passenger Lists (1820 - 1914)*

We construct our immigration database from two primary sources: the Castle Garden Immigration Database and the Hamburg Passenger Lists.

The *Castle Garden Immigration Database (1820-1892)* is an educational project of the Battery Conservancy. The database contains the list of all immigrants entering the US via the port of New York between 1820 to 1892. In total, the database comprises approximately 11 million individual micro-records. Through a cooperation with the Battery Conservancy, we have access to the entire Castle Garden Immigration database. Importantly, these immigration records contain detailed pre-immigration occupation information of respective immigrants.

We complement this database with original passenger lists of all immigrants leaving from the port of Hamburg to US between 1850 and 1914 called the *Hamburg Passenger List Database (1850-1914).* We have access to the complete records through a cooperation with the Hamburg State Archive. For our main analysis, we translate these passenger lists with pre-immigration occupation in German to English and construct a panel data of immigrants leaving from the port of Hamburg to US demographic census records by using time-invariant information such as name, gender, year of birth, arrival year and place of birth. Details regarding record linking are discussed in 2.3.2.

2.3.1.2 *The Complete Count Federal Demographic Decennial Census (1850-1940)*

To measure the impact of immigration, we require information on immigrants' characteristics after their arrival in the US. We do so by linking the individual immigration records from the Immigration Database with the Complete Count Federal Demographic Census. We take advantage the complete transcription of Federal Census Records between 1850 to 1940. The US federal demographic census year records exist for all years from 1850 to 1940 every decade except for 1890 (which was lost due to fire)

Our procedure of linking the immigration data to the US census data, which we describe in detail in Section 2.3.2, allows us to construct a unique dataset, where we can observe both the pre-migration occupation and the entire employment lifecycle in the US for millions of immigrants during the study period. In terms of observables, the complete-count Demographic Census contains various socio-economic characteristics. In particular, occupation classification following the *Historical International Classification of Occupations (OCCHISCO)* is extremely detailed and well-suited to measure immigration induced knowledge flows.

2.3.1.3 *Measuring Productivity: The Historical Census of Manufacturers (1870 - 1929)*

Our empirical strategy heavily relies on spatial variation in productivity growth, innovation activity and the settlement of immigrants. Measuring productivity at a fine spatial resolution in the 19th Century is difficult. First of all, there are no measures of wages at the county-level. While the available individual-level data in the decennial Population Census contains county-identifiers, earnings have only been reported starting in 1940. Secondly, information on labor earnings stemming from the National Accounts is available in the 19th Century, but the data does not have a spatial dimension. To overcome this problem, we digitize the published results from the Census of Manufacturers. These tables are published at either the county-industry level or city-industry level and report standard information from firms balance sheets. In particular, they report the number of manufacturing establishments, total number of workers, value of manufactured output, wage-bill and value of the capital stock. As an example, recall the information on the Steinway & Son Factory shown in Figure 23. As our main measures of spatial productivity, we consider total value added per manufacturing employee, manufacturing value added relative to the wage bill or manufacturing revenue TFP, i.e. $Y_r/\left(K_r^{\alpha}L_r^{1-\alpha}\right)$, where Y_r is total value added and K_r and L_r denote the capital stock and employment. We digitize this data at the *county*-industry level for the years 1860, 1870, 1880 and 1929 and at the *city*-industry level for the years 1880, 1900, 1909, 1919, 1929 and 1939. In Section below, we explain how we harmonize this information and combine with other sources.

2.3.1.4 *Measuring Innovative Activity: Data on Patenting*

To measure the extent of innovative activity at the county level, we exploit information on patenting. The United States Patent and Trademark Office (USPTO) granted millions of patents since 1790 and all patents contain information about the location of the patent and can therefore be geo-referenced. While this information is publicly available in the "HistPat Dataset", we need to harmonize the information on patents at the level of standard industrial classifications.

We use the information on patenting in two ways. We first use the extent of patenting in a given region-sector cell as a measure of idea creation. This measure is consistent with a simple idea-based growth model and can easily be mapped to productivity growth in a model-consistent way.[5] More importantly, we use the patent data to devise a novel measure of *spatial idea novelty.* Our theoretical mechanism stresses the importance of European immigrants bringing *novel* insights to the US. To relate immigration inflows to the novelty of the type of ideas invented in a particular county, we use textual analysis to measure the extent to which patents originating in a particular region are similar to the patents that have been invented in that region in the past. Intuitively, as in Kelly et al. (2018), we measure the similarity between two patents as the correlation in the words, respective patents use. Given this measure of patent-to-patent similarity, we then calculate the similarity of ideas invented in a region as the average patent-to-patent similarity between new patents originated in a particular region and the set of patents stemming from a particular region in the past. Our measure of *spatial idea novelty* is then simply the inverse of this average similarity and we can relate it systematically to the spatial inflow of immigrants.

[5] Many contributions using more recent data on patenting stress the importance of weighing patents by their subsequent citations to adjust patents for their quality. Such quality adjustments are not possible in the 19th century patent data as the information on citations patterns is too scarce.

2.3.2 Dataset Construction: Record Matching And Harmonization

To combine the four datasets introduced in Section 2.3.1 into a single database, we need to (i) link the micro-level Immigration Database with the Full Count US Demographic Census and (ii) provide consistent crosswalks to merge patents, measures of productivity and immigrant inflows at the spatial and sectoral level. In this section we provide more details for these two steps.

2.3.2.1 Record Matching: Linking Immigrants and the Complete Count Federal Demographic Decennial Census, 1850-1940

We have access to the complete individual-level complete-count US demographic federal census records from 1850-1940.[6] By linking our novel immigration records to the US census records, we can measure the entire life-cycle of immigrants since they entered the US.

Our record linking procedure has the following characteristics: (1) we rely on all *complete-count* US Federal Census records with *occupation and industry information* (a newly transcribed variable from the original census records), (2) we link people at *more than two points in time* and (3) we use both *individual and household level information* to improve the matching of individuals.

These three elements are important for this study. The availability of the occupation information both before and after entering the US enables us to measure individuals' skills and to investigate novel economic problems, such as occupational transitions along immigrants' life-cycle. Also, while many record matching methods match individuals only at two points in time, this horizon may not be long enough to systematically analyze spatial mobil-

[6] We use restricted complete-count US demographic census from 1850 to 1940 to link individual-level records by implementing random forest classification. This record linking methodology is similar in spirit to Minnesota Population Center (MPC) Record record linkage project of the 1850-1930 sample census records to the 1880 complete count census records. However, MPC implemented a support vector machines (SVM) to automate the record linking. Feigenbaum (2016) discusses a machine learning approach to census record linking and compares different possible matching algorithms. He proposes a probit-based method as an ideal choice.

Year	Record Type	Germany	Italy	Ireland	UK	Total
1880	Immigration records	1,530,063	10,843	732,100	392,528	2,665,534
	Census record	1,081,249	31,292	887,609	514,476	2,514,626
	Matches	65,962	543	21,823	26,305	114,633
1900	Immigration records	2,543,625	616,566	761,317	438,249	4,359,757
	Census record	1,431,253	315,105	746,408	639,007	3,131,773
	Matches	309,198	48,527	62,119	47,790	467,634
1910	Immigration records	2,528,249	852,445	761,317	438,251	4,580,262
	Census record	1,343,333	887,044	615,329	676,429	3,522,135
	Matches	281,619	88,703	45,571	40,611	456,504
1920	Immigration records	2,384,298	859,328	713,603	400,094	4,357,323
	Census record	862,070	955,014	460,635	605,809	2,883,528
	Matches	194,219	75,086	25,707	27,590	322,602

Notes: The table contains the total number of immigrants (by county of origin) from the Castle Garden immigration records that migrated to the US *before* the respective year ("Immigration records"), the number of immigrants (by county of origin) in the US Population Census and the number of matches. The last column contains the sum of the four origin countries.

Table 21: Linking Immigration Records and the US Population Census

ity along the life-cycle. In this project we match individuals multiple points in time (up to three or four times) so that we follow individuals for multiple decades. Finally, most existing historical record matching practices drop non-unique potential matches. This may introduce a systematic bias of matched records, as for example records with relatively common names will be systematically excluded. We therefore match and analyze the characteristics of such non-unique matches by gender, race, and birthplace group and have developed new methods to find unique matches by using additional household-level information.

The current results of our record matching procedure are contained in Table 21. The results are preliminary as we have currently only matched four major immigrants groups by sending country (Germany, Italy, Ireland, and the UK). Although not included in the table, we also matched the records for the years 1850 and 1930. As in Table 21, our procedure yields millions of matched immigrant records which we can track their lives for multiple decades. We have already matched more than a half million immigrant records from four major immigrant sending countries between 1850 and 1930.

2.3.2.2 Record Matching: Linking US Patent Records and the Complete Count Federal Demographic Decennial Census, 1850-1940

We link the patent-level US Patent Records from the US Patent and Trademark Office to historical US Census to investigate the characteristics of inventors in the US. In explaining the record linking procedure, we will take an example of Year 1910 as a reference here. The record linking algorithm proceeds as follows: consider the universe of patents published pre-1910. Each of these patents has (via HistPat) an associated county and (most, but not all, have) an associated inventor name. Using individual-level information such as first and last name, and geographic information such as state and county, consider the universe of individuals who, as of the 1910 census, were residing in that county. This forces the assumption that individuals do not move after patenting. Then, we measure the Jaro-Winkler distance between the first name of patenters and the listed first name of individuals (males) in the census, and similarly for the last name for string comparisons.

2.3.2.3 Data Harmonization and Integration at the Country-Industry Level

This project relies on a micro-to-macro approach, where we combine the microdata from the matched Population Census with additional data sets at the macro (i.e. county-industry) level. To do so, it is essential to harmonize the different data sources in a unified format.

As the occupation information in the *Hamburg Passenger Lists* in only available in German and they are neither integrated nor coded with any standardized classification system (recall Mr Steinweg's occupation "Instrumentenmacher" in Figure 21), we translate the reported German occupational string and codify them in the occupational system of the US Federal Census. Similarly, the occupational classification available in the *Castle Garden Database*, while available in English, is also not classified. Therefore, we construct occupational crosswalk between Castle Garden occupation strings and occupational measures which are consistently available in US demographic census (i.e. variaables "OCCHISCO" and "OCC1950").

Similarly, we match the reported sectoral groups in the Census of Manufacturers to a standard classification of industries. As seen in Figure 23, our Manufacturing Census data is more detailed than 1950 Census Bureau industrial classification system which is available in the US census. Therefore, we create industry crosswalk from reported industry strings in the Census of Manufacturers to 1950 Census Bureau industrial classification. The same applies to the historical patent data, where we use a combination of the detailed patent descriptions and the patent classifications to assign individuals patents to particular sectors of production.

Finally, we merge all aforementioned datasets at a unified spatial level. While the Population Census and the Patent Data is already geo-referenced, we perform geo-referencing using the original geographical information in the Manufacturing Census. We integrate all county aggregates from 1860 to 1940 at a consistent level of aggregation. As counties and state boundaries have changed over time, we take the county shapefiles from National Historical Geographic Information System, and create the geographic location-based crosswalks of counties across time.

2.4 Setup

The country is divided in R regions denoted by r. Time is discrete. We assume the existence of iceberg trade costs between locations. Thus, the final good price of goods from location r in location j is given by

$$p_{rjt} = \tau_{rj} p_{rt},$$

where p_{rt} denotes the price in location r. All workers have identical Constant Elasticity of Substitution (CES) preferences over goods of all locations r arriving in region j, c_{rjt}, at period t given by

$$C_{jt} = \left(\sum_r c_{rjt}^{\frac{\varepsilon-1}{\varepsilon}} \right)^{\frac{\varepsilon}{\varepsilon-1}}, \tag{2.1}$$

where ε is the elasticity of substitution across varieties.

Total spending (and income) of the representative agent in region r is denoted by E_{rt}. Workers can work as production and research workers with associated wages in location r, w_r^P and w_r^R. The native population in each region is denoted by L_{rt} and the number of immigrants is denoted by I_{rt}.

Each region produces a final tradable good, which we denote by Y_{jt}. The production of this final good in each location requires a unit continuum of differentiated, non-tradable varieties i, $x_t(i)$, so that

$$Y_{rt} = Z_{rt}^A \left(\int_{i=0}^{1} x_t(i)^{\frac{\sigma-1}{\sigma}} \, di \right)^{\frac{\sigma}{\sigma-1}}, \tag{2.2}$$

where Z_{rt} is a regional exogenous productivity term. The price of the good produced in in the region is given by

$$p_{rt} = \frac{U_{rt}}{Z_{rt}^A},$$

where U_{rt} denotes the cost of production of the intermediate input bundle and is given by

$$U_{rt} = \left(\int_{i=0}^{1} u_{rt}(i)^{1-\sigma} \, di \right)^{\frac{1}{1-\sigma}},$$

where $u_{rt}(i)$ is the price per unit of variety i in region r at time t.

2.4.1 Firms and Innovation

Firms are monopolists for their differentiated varieties. Firms differ by location and efficiency. The production function for varieties in region r with efficiency z is given by

$$x_{rt}(z) = z \left(q_{rt}(z) \right)^{\frac{1}{\sigma-1}} h_{rt}(z) \tag{2.3}$$

where h_{rt} denotes the total amount of efficiency units hired for production by a firm in region r time t, q_{rt} denotes the quality of firm z in region r and z is an exogenous, firm-specific efficiency.[7]

While z is an exogenous firm-characteristic, which is constant, quality q evolves endogenously. In particular, firms can increase their current quality $q_{rt}(z)$ by a factor $i_{rt}(z)$ according to

[7] The scaling $q^{\frac{1}{\sigma-1}}$ is a normalization that allows to write profits in a linear form, ultimately.

$$q_{rt+1}\left(z\right) = q_{rt}\left(z\right) i_{rt}\left(z\right). \tag{2.4}$$

Note that we, in principle, allow for depreciation of the firm quality i.e. if firms do not innovate enough, their productivity might decline, $i_t < 1$. As we show below, this will not be the case along a Spatial Balanced Growth Path. To innovate, firms need to hire researchers. More specifically, we assume that to increase their quality by $i_{rt}\left(z\right)$, firms pay a cost of

$$c^I_{rt}\left(q; z\right) = \frac{1}{\zeta_r Z^I_t} \frac{z^{\sigma-1} q}{Q^{\lambda}_{rt}} \frac{i_t^{\iota}}{\iota} w^R_{rt}, \tag{2.5}$$

where Q_{rt} is the average productivity in region r at time t

$$Q_{rt} = \mathbb{E}_{rt}\left[z^{\sigma-1} q\left(z\right)\right] = \int z^{\sigma-1} q_{rt}\left(z\right) dF_{rt}\left(z\right) \tag{2.6}$$

and $F_{rt}\left(z\right)$ is the cross-sectional distribution of firm efficiencies in region r at time t. Equation (2.5) stresses that the costs of innovation depend on both firm-level and regional characteristics. On the regional level, they are determined by the prevailing research wage in location r, w^R_{rt}, a fixed region-specific "innovation efficiency" ζ_r and a spill-over term Q^{λ}_{rt}. The parameter λ governs the extent to which research cost fall in the existing level of productivity Q_{rt}. As we show below, if $\lambda = 1$ the model becomes an endogenous growth model and if $\lambda < 1$ the model is a semi-endogenous growth model. We also show that this distinction makes precise predictions on the long-run effect of immigrant inflows on regional economic activity in the long-run. As in Atkeson and Burstein (2010) we also assume that innovation costs are linear in firm efficiency, $z^{\sigma-1}q$. This scaling implies that the model is consistent with Gibrat's Law where growth is independent of size. Finally, Z^I_t is a time-varying efficiency shifter determining the cost of innovation, which is common across all locations.

Firm optimization The constant elasticity aggregator across intermediate input producers implies that firms' prices are given by a constant markup over the production cost

$$u_{rt}\left(z\right) = \frac{\sigma}{\sigma-1} \frac{w^P_{rt}}{z q_{rt}\left(z\right)^{\frac{1}{\sigma-1}}}. \tag{2.7}$$

Here w_{rt}^{P} is the wage for production workers in region r at time t. Standard arguments imply that firm z in region r at time t has a variable profit of

$$\pi_{rt}\left(z\right)=\frac{1}{\sigma}\left(\frac{u_{rt}\left(z\right)}{U_{rt}}\right)^{1-\sigma}E_{rt}=\frac{1}{\sigma}\frac{z^{\sigma-1}q\left(z\right)}{Q_{rt}}E_{rt}. \tag{2.8}$$

Firms' innovation decisions are of course dynamic in nature. Letting the real interest rate be r_t, the value function of a firm in region r at time t with a productivity $z^{\sigma-1}q\left(z\right)$ is given

$$V_{rt}\left(z^{\sigma-1}q\right)=\pi_{rt}\left(z^{\sigma-1}q\right)+\max_{i_t}\left[\frac{1}{1+r_t}V_{rt+1}\left(z^{\sigma-1}qi_t\right)-\frac{1}{\zeta_r Z_t^I}\frac{z^{\sigma-1}q}{Q_{rt}^{\lambda}}\frac{i_t^{\iota}}{\iota}w_{rt}^{R}\right]. \tag{2.9}$$

Hence, $\frac{1}{1+r_t}V_{rt+1}\left(z^{\sigma-1}qi_t\right)$ is the expected value of being a firm with productivity qi_t in period $t+1$. Consider the value function $V_{rt}\left(q\right)$ in (2.9). The value function is linear homogeneous in q, i.e. $V_{rt}\left(z^{\sigma-1}q\right)=z^{\sigma-1}qv_{rt}$, where

$$v_{rt}=\frac{1}{\sigma}\frac{E_{rt}}{Q_{rt}}+\frac{\iota-1}{\iota}\frac{w_{rt}^{R}}{\zeta_r Z_t^I}\frac{i_t^{\iota}}{Q_{rt}^{\lambda}}, \tag{2.10}$$

and the optimal rate of innovation i_{rt} is given by

$$i_t=\left(\frac{v_{rt+1}}{1+r_t}\frac{Q_{rt}^{\lambda}\zeta_r Z_t^I}{w_{rt}^{R}}\right)^{\frac{1}{\iota-1}}. \tag{2.11}$$

Proof. See Section B.1.1 in the Appendix. □

Proposition 2.4.1 shows that innovation incentives i_{rt} are equalized across all firms in location r. This is an implication of the homogeneity of the value function. The policy function for firms' innovation incentives in (2.11) shows that the optimal innovation rate depends on the discounted future value $\frac{v_{rt+1}}{1+r_t}$ relative to the cost of innovation $\frac{Q_{rt}^{\lambda}\zeta_r Z_t^I}{w_{rt}^{R}}$. Note also that by combining (2.10) and (2.11) we can express the value function as a forward looking difference equation

$$v_{rt}=\frac{1}{\sigma}\frac{E_{rt}}{Q_{rt}}+\left(\frac{\iota-1}{\iota}\right)\left(\frac{Q_{rt}^{\lambda}\zeta_r Z_t^I}{w_{rt}^{R}}\right)^{\frac{1}{\iota-1}}\left(\frac{v_{rt+1}}{1+r_t}\right)^{\frac{\iota}{\iota-1}}. \tag{2.12}$$

2.4.2 Labor Supply

In our model, individuals have two margins for their labor supply decisions: they decide which sector to work in and which location to migrate to. In terms of timing, we assume that individuals first decide on their geographical location r and then on their preferred sector of employment.

Labor supply across sectors We model sectoral labor supply with a simple Roy structure. Individuals are characterized by a single attribute - their immigration status, which we denote by $n \in \{N, I\}$, where $n = N$ denotes “Natives”and $n = I$ denotes “Immigrants”. We assume that individual i draws a vector of efficiency

$$\{x^P, x^R \quad ,$$

where x^P and x^R denotes the efficiency units as a production worker and a research worker. We assume that x^P and x^R are drawn independently from the following Frechet distribution

$$F_{jn}(x) = e^{-h_n^j x^{-\theta}}. \tag{2.13}$$

Here h_n^j parametrizes the average human capital of an individual with nationality $n$in sector $j \in \{R, P\}$ and θ parametrizes the labor supply elasticity.

Standard arguments imply that the share of people of type n working in sector $j = R, P$ in region r is given by

$$s_{rnt}^j = \frac{h_n^j \left(w_{rt}^j\right)^\theta}{h_n^P \left(w_{rt}^P\right)^\theta + h_n^R \left(w_{rt}^R\right)^\theta}. \tag{2.14}$$

Similarly, the aggregate level of human capital provided by workers of type n in region r towards sector $j \in \{R, P\}$ is given by

$$H_{rnt}^j = L_{rnt}\Gamma\left(1 - \frac{1}{\theta}\right)\left(h_n^j\right)^{\frac{1}{\theta}}\left(s_{rnt}^j\right)^{\frac{\theta-1}{\theta}}, \tag{2.15}$$

where L_{rnt} is the number of people of type n in region r. Hence, the aggregate supply of efficiency units provided to sector j in region r is given by

$$H^j_{rt} = L_{rt}\Gamma\left(1-\frac{1}{\theta}\right)\left[\left(1-\varpi_{rI}\right)\left(h^j_N\right)^{\frac{1}{\theta}}\left(s^j_{rNt}\right)^{\frac{\theta-1}{\theta}} + \varpi_{rI}\left(h^j_I\right)^{\frac{1}{\theta}}\left(s^j_{rIt}\right)^{\frac{\theta-1}{\theta}}\right] \; j\in\{I,N\}, \tag{2.16}$$

where $\varpi_{rI} = L_{rI}/L_r$ is the population share of immigrants in region r. Note that the respective employment shares s^j_{rN} only depend on relative wages (see (2.14)). Hence, the aggregate supply of human capital towards the research and the production sector also only depends on the relative wage within region r.[8]

For future reference we distinguish two special cases as delineated in Arkolakis et al. (2018). If skills are inelastically provided (which corresponds to the case of $\theta \to 1$), the share of people of group n working in sector j and the total amount of human capital is given by

$$s^j_{rnt} = \frac{h^j_n}{h^R_n + h^P_n} \text{ and } H^j_{rtn} = L_{rtn}h^j_n.$$

The second polar case is the case of homogeneity, i.e. $\theta \to \infty$. In that case, workers' occupation choice problem takes a simple cutoff-rule

$$\text{work in sector } R \text{ if and only } h^R_{rtn}w^R_{rt} \geq h^P_{rtn}w^P_{rt}.$$

If, for example $h^R_{rtn} = h^P_{rtn} = 1$, and with an interior solution where all locations produce and innovate we get that wages are equalized across sectors, i.e.

$$w_{rt} = w^R_{rt} = w^R_{rt}$$

and that the share of people working in the two sectors are fully demand determined.[9]

[8] Note that if natives and immigrants are identical, i.e. $h^j_N = h^j_I$, (2.16) reduces to the usual expression $H^j_r = L_r\Gamma\left(1-\frac{1}{\theta}\right)\left(h^j\right)^{\frac{1}{\theta}}\left(s^j_r\right)^{\frac{\theta-1}{\theta}}$, as sectoral employment shares will be equalized given that they face the same prices.

[9] Note that if natives and immigrants differ in h^j_{rtn}, generically one group will be fully specialized.

2.4.3 Aggregation

Above we have characterized the optimal decisions of firms and workers. To calculate the equilibrium (in particular to solve for equilibrium prices), we need to aggregate these decisions. Exploiting the functional form assumptions of our setup we can characterize the aggregate law of motion for average quality Q_{rt} and express aggregate output directly in terms of production workers' labor supply.

An Aggregate Production Function Our economy aggregates - at the production side - to a standard macro-spatial model. In particular, we can define an aggregate production function of each location that is linear on the total supply of efficiency units of labor in region r and time t and its aggregate productivity is determined by aggregate object of the economy. This result is formalized in the following Lemma Consider the model above. Let H^P_{rt} be the total supply of efficiency units of production workers in region r at time t. Aggregate output of the tradable good in region r is given by

$$Y_{rt} = A_{rt} \times H^P_{rt},$$

where the endogenous TFP term A_{rt} is given by

$$A_{rt} = Z^A_{rt} \left(Q_{rt}\right)^{\frac{1}{\sigma-1}} . \tag{2.17}$$

Proof. See Section B.1.3 in the Appendix. □

Lemma 2.4.3 and (2.23) show that the evolution of spatial labor supply $\left\{H^P_{rt}, H^R_{rt}\right\}$ fully summarize the evolution of aggregate output $\{Y_{rt}\}$: $\left\{H^R_{rt}\right\}$ determines the evolution of productivity $\{Q_{rt}\}$ from (2.23) and $\left\{H^P_{rt}\right\}$ determines aggregate output as in a standard model of trade.

Aggregate Trade Between Regions Regional trade flows are statically determined by considering firm prices. Each consumer consumes products from different locations. Given the assumption of CES demand, the market share of location r in the basket of location j at time t is given by

$$\lambda_{rjt} = \frac{\int \left(u_{rt}\left(z\right) dF_{rt}\left(z\right)\right)^{1-\sigma}}{\sum_{r'} \int \left(u_{r't}\left(z\right) dF_{r't}\left(z\right)\right)^{1-\sigma}} = \frac{\left(\frac{w^P_{rt}}{A_{rt}}\right)^{1-\sigma} \int z^{\sigma-1} q\left(z\right) dF_{rt}\left(z\right)}{\sum_{r'} \left(\frac{w^P_{r't}}{A_{r't}}\right)^{1-\sigma} \int z^{\sigma-1} q_{r'}\left(z\right) dF_{r't}\left(z\right)}. \tag{2.18}$$

Moreover, one can show that production workers receive a constant share of aggregate income. Hence, total spending, E_{rt}, can be written as a function of production worker income

$$E_{rt} \equiv w^P_{rt} H^P_{rt} \frac{\sigma}{\sigma - 1}. \tag{2.19}$$

We assume that trade is balanced period-by-period, i.e. that product markets clear every period. In other words we have that in equilibrium

$$E_{rt} = \sum_j \lambda_{rjt} E_{jt}. \tag{2.20}$$

Aggregate labor demand

The market for labor for innovation implies that the total number of efficiency units of workers in innovation must be equal to

$$H^R_{rt} = \frac{1}{\iota} \frac{1}{\zeta_r Z^I_t} \frac{\int z^{\sigma-1} q\left(z\right) dF_{rt}\left(z\right)}{Q^\lambda_{rt}} i^\iota_t = \frac{1}{\iota} \frac{i^\iota_t}{\zeta_r Z^I_t} \frac{1}{Q^{\lambda-1}_{rt}} \tag{2.21}$$

Combining this equation with the first order condition for innovation equation (2.10) we obtain the labor demand for innovation

$$H^R_{rt} = \frac{1}{\iota} \left(\frac{v_{rt+1}}{1 + r_t} \frac{1}{w^R_{rt}} \right)^{\frac{\iota}{\iota-1}} \left(Q^\lambda_{rt} \zeta_r Z^I_t\right)^{\frac{1}{\iota-1}} Q_{rt}. \tag{2.22}$$

The Endogenous Law of Motion for Aggregate Quality

Using equation (2.4) and aggregating among firms given the innovation equation (2.10) we obtain

$$Q_{rt+1} = Q_{rt} i_{rt}$$

Combining this equation, with (2.22) we can directly relate the the demand for research workers to the resulting growth rate

$$\frac{Q_{rt+1}}{Q_{rt}} = \left(H^R_{rt}\zeta_r Z^I_t Q^{\lambda-1}_{rt}\iota\right)^{1/\iota}, \tag{2.23}$$

i.e. the evolution of local productivity Q_{rt} is fully determined from the equilibrium amount of researchers H^R_{rt}.

We exploit these aggregation results to define the equilibrium as a macro system of discrete blocks of equations, statics and dynamic.

2.4.4 Dynamical Equilibrium

To characterize the equilibrium of the model we need to consider the market clearing of both product and labor markets. 2.22 *A dynamical equilibrium are sequences of production and research wages* $\left\{w^P_{rt}, w^R_{rt}\right\}_{rt}$*, per-quality-unit value functions* $\{v_{rt}\}_{rt}$*, innovation choices* $\{i_{rt}\}_{rt}$*, labor allocations* $\left\{H^P_{rt}, H^R_{rt}\right\}_{rt}$*, regional qualities* $\{Q_{rt}\}_{rt}$*, and consumption demands* $\{c_{jrt}\}_{jrt}$ *such that given an initial level of regional quality* $\{Q_{r0}\}_r$*, labor and good markets clear at each point time,*

1. *firms' innovation choices* $\{i_{rt}\}_{rt}$ *are consistent with* $\{v_{rt}\}_{rt}$*, i.e. solve (2.11)*
2. *the evolution of qualities* $\{Q_{rt}\}_{rt}$ *is consistent with firms' innovation choices* $\{i_{rt}\}_{rt}$*, i.e. solve (2.23),*
3. *the per-quality-unit value functions* $\{v_{rt}\}_{rt}$ *solve (2.10).*
4. Labor markets clear, i.e. labor demand and supply for production and research labor equalize given equations (2.22), (2.20), and (2.16).

Definition 2.4.4 makes clear that the general equilibrium consists of a set of dynamic and static equations for an arbitrary number of regions. To solve this daunting dynamic fixed point problem we follow a strategy of modularization, as in Adao et al. (2019), in order to determine sets of equations that can be independently solved taking a subset of the variables of the system at a time. The difference with our approach is that one of the equations of the paper, the firm-innovation module described below, contains dynamic difference equations and not just static. The three modules are as follows:

1. *The Labor Module:* Given production worker wages $\left\{w^P_{rt}\right._{r}$, average product quality $\{Q_{rt}\}_r$ and the value of innovation $\{v_{rt+1}\}_r$, we can use the labor supply equations 2.16 and the labor demand for research workers (2.22) to determine $\{H^P_{rt}, H^R_{rt}, w^R_{rt}\}_r$,

2. *The Trade Module:* Given production worker labor supply $\left\{H^P_{rt}\right._{r}$ and average product quality $\{Q_{rt}\}_r$, production wages $\left\{w^P_{rt}\right._{r}$ are determined from the goods markets. Using (2.20) we obtain that

$$w^P_{rt}H^P_{rt} = \sum_j \frac{\left(\frac{w^P_r}{A_{rt}}\tau_{rj}\right)^{1-\varepsilon}}{\sum_r \left(\frac{w^P_r}{A_{rt}}\tau_{rj}\right)^{1-\varepsilon}} w^P_{jt}H^P_{jt}. \tag{2.24}$$

3. *The Innovation Module: Given wages and labor allocations* $\left\{H^P_{rt}, H^R_{rt}, w^P_{rt}, w^R_{rt}\right._{r}$ and the level of quality $\{Q_{rt}\}_r$, we can solve for i_{rt} from (2.23). Given i_{rt} we can solve for $\{v_{rt}, v_{rt+1}\}$ from *(2.10) and (2.11).*

2.4.5 Innovation in Space and Time

Starting from the value function, equation 2.12,

$$v_{rt} = \frac{1}{\sigma}\frac{E_{rt}}{Q_{rt}} + \left(\frac{\iota-1}{\iota}\right)\left(\frac{Q^\lambda_{rt}\zeta_r Z^I_t}{w^R_{rt}}\right)^{\frac{1}{\iota-1}}\left(\frac{v_{rt+1}}{1+r_t}\right)^{\frac{\iota}{\iota-1}}.$$

Now use 2.23, and $E_{rt} = \frac{\sigma}{\sigma-1}w^P_{rt}H^P_{rt}$ to obtain

$$v_{rt} = \frac{1}{\sigma-1}\frac{w^P_{rt}H^P_{rt}}{Q_{rt}} + \left(\frac{\iota-1}{\iota}\right)\iota^{\frac{1}{\iota-1}}\left(\frac{Q_{rt+1}}{Q_{rt}}\right)(Q_{rt+1})^{\frac{1}{\iota-1}}\frac{v_{rt+1}}{1+r_t} \tag{2.25}$$

We now analyze aspecial case of our model, aspecific factor case, where workers provide their skills inelastically to the two production sectors.

Thecase of no mobility across sectors that corresponds to $\kappa \to 1$ where $H_{rt}^P = \bar{H}_{rt}^P$, $H_{rt}^R = \bar{H}_{rt}^R$ (see Arkolakis et al. (2018)).

Notice that in the second case equation (2.23) implies that the law of motion of Q_{rt} does not depend on anything else other than $\zeta_r Z_t^I$. In other words, the law of motion of Q_{rt} can be completely determined by this equation and it is country-by-country specific.

Assuming that indeed $\zeta_r Z_t^I$ are fixed and we start from a steady state $Q_{rt} = \bar{Q}_r$. Then using (2.25) the value function of the firm is given by

$$v_{rt} = \frac{1}{\sigma - 1} \frac{w_{rt}^P \bar{H}_{rt}^P}{\bar{Q}_r} + \left(\frac{\iota - 1}{\iota} \right) \iota^{\frac{1}{\iota - 1}} \frac{v_{rt+1}}{1 + r_t}.$$

The solution of the system is then vectors of wages and value function $\{v_{rt}\}, \{w_{rt}^P\}$ that solve the above equation and

$$w_{rt}^P \bar{H}_{rt}^P = \sum_j \frac{\left(\frac{w_{rt}^P}{A_{rt}} \tau_{rj} \right)^{1-\varepsilon}}{\sum_r \left(\frac{w_{rt}^P}{A_{rt}} \tau_{rj} \right)^{1-\varepsilon}} w_{jt}^P \bar{H}_{jt}^P,$$

with $A_{rt} = Z_{rt}^A \left(\bar{Q}_{rt} \right)^{\frac{1}{\sigma - 1}}$.

2.5 The Spatial Balanced Growth Path

We now characterize the balanced growth path (BGP) of this economy. Along the BGP, the distribution of wages and spending across regions is stationary. This requires that productivity grows at the same rate in all regions. The characterization of the BGP is contained in the following proposition. Consider the economy above and consider a BGP. Along the BGP, wages $\left\{ w_{rt}^P, w_{rt}^R \right\}_r$ and spending $\{E_{rt}\}_r$ grow at the rate of TFP A_{rt} and the population distribution is stationary.

1. The growth rate of TFP A_{rt} is constant across regions and is given by

$$1 + g_A = \left(1 + \overline{g}_Z \right) \left(1 + \overline{g}_M \right)^{\frac{1}{1-\lambda} \frac{1}{\sigma - 1}},$$

where $\overline{g}_Z$ is the exogenous growth rate of Z_{rt} and $\overline{g}_M$ is the exogenous growth rate of research productivity M_t.

2. The growth rate of productivity Q_{rt} is given by

$$1 + g_Q = (1 + \overline{g}_M)^{\frac{1}{1-\lambda}}$$

3. The value function v_{rt} is given by

$$v_{rt} = \frac{1}{1 - \frac{\iota - 1}{\iota} \frac{1+g_A}{1+r}} \frac{1}{\sigma} \frac{E_{rt}}{Q_{rt}},$$

   where E_{rt} is total spending on goods in region r

4. The spending share on production workers and researchers is equalized across space, i.e.

$$\frac{H_{rt}^R w_{rt}^R}{H_{rt}^P w_{rt}^P} = \frac{1}{\sigma - 1} \frac{\frac{1}{\iota} \frac{1+g_A}{1+r}}{1 - \frac{\iota - 1}{\iota} \frac{1+g_A}{1+r}}. \tag{2.26}$$

5. The distribution of productivity across space satisfies

$$\frac{Q_{rt}}{Q_{jt}} = \left. \frac{\zeta_r}{\zeta_j} \times \frac{\frac{E_{rt}}{w_{rt}^R}}{\frac{E_{rt}}{w_{rt}^R}} \right)^{\frac{1}{1-\lambda}} \quad \text{for all } r, j \tag{2.27}$$

Proof. See Section B.1.4 in the Appendix. □

The main implication of Proposition 2.5 is contained in (2.27): the long run distribution of productivity across space is endogenously determined. Using the fact that $E_{rt} \propto H_{rt}^P w_{rt}^P$ (see (2.19)) and $H_{rt}^P w_{rt}^P \propto H_{rt}^R w_{rt}^R$ (see (2.26)), we can express (2.27) as

$$\ln\left(\frac{Q_r}{Q_j}\right) = \underbrace{\frac{1}{\lambda} \ln\left(\frac{\zeta_r}{\zeta_j}\right)}_{\text{Exogenous differences in research efficiency}} + \underbrace{\frac{1}{\lambda} \ln\left(\frac{H_{rt}^R}{H_{jt}^R}\right)}_{\text{Supply of researchers}}. \tag{2.28}$$

Hence, region r has high productivity relative to region j if it is relatively efficient to produce new ideas, i.e. $\zeta_r > \zeta_j$, and it is able to attract relatively more researchers, i.e. $H_{rt}^R > H_{jt}^R$. The relative abundance of researchers is of course endogenous and determined from both the trade equilibrium and the system of equations governing labor mobility. Moving costs or the degree of openness across regions as part of the trade module will therefore affect the long-run distribution of productivity across space.

2.6 Reduced form evidence

In this section, we provide direct evidence for the mechanism in our model. In particular, we show that immigration inflows are positively related to regional productivity growth and to patent activity.

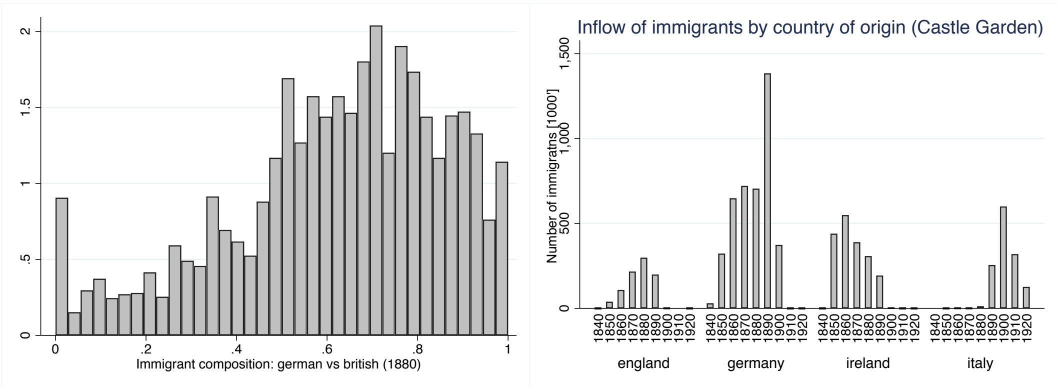
2.6.1 Constructing an instrument for the allocation of immigrants

In Figure 25 we show the conceptual idea for our instrument, which is based on CARD. We construct predicted immigrant flows from the time-series of the aggregate inflow of immigrants from different countries origin interacted with the existing cross-sectional distribution of immigrants prior to our sample. The time-series variation is shown in the right panel of Figure 25. Two aspects are interesting. First of all, there is a substantial time-series variation *within* immigrant groups. While the flow of Irish immigrants is declining after 1860, the immigration from Germany is increasing and has a peak in the decade between 1880 and 1890. At the same time there is large cross-sectional variation in the composition of the immigrant population across counties in the US. As an example we depict the share of german immigrants relative to immigrants from the UK across US counties in 1880 in the left panel of Figure 25. It is clearly seen that the cross-sectional variation is large. While many counties have a large "surplus" in German immigrants relative to British immigrants other counties are much more populated by british immigrants relative to Germans.

Given these two sources of variations, we construct a *predicted* immigrant stock in county r as follows. Let I_{r1860} be the total immigrants in county r in 1860. Also, let I^n_{r1860} denote the total number of immigrants with nationality n in county r in 1860. We then construct the predicted stock of immigrants in county r for all $t > 1860$, I^P_{rt}, as

$$I^P_{rt} = I^P_{rt-1} + \sum_{n \in \{GER,UK,ITA,IRL\}} IF^n_t \times \frac{I^n_{r1860}}{\sum_r I^n_{r1860}}, \tag{2.29}$$

where IF^n_t is the aggregate inflow of immigrants in year t as depicted in Figure 25. Intuitively, we assign the inflowing immigrants from country n according to their cross-sectional



Note: The left panel shows the cross-sectional distribution $\frac{I_r^{GER}}{I_r^{GER}+I_r^{UK}}$, where I_r^{GER} (I_r^{UK}) is the number of german (english) immigrants living in region r in the year 1880. In the right panel we display the time-series of immigration inflows by decade for the four main countries of origin.

Figure 25: Constructing the Instrument

distribution in 1860 and then accumulate the regional immigrant stock with these inflow. As we only use (2.29) as an instrument, we abstract from mortality, which in principle might be specific to particular regions and nationalities.

Because many of our regression utilize the share of immigrants in a particular locality as a regressor, we also construct a predicted immigrant share as

$$s_{Irt}^P = \frac{I_{rt}^P}{I_{rt}^P + L_{Nrt}},$$

where L_{Nrt} is the native population in region r at time t.

2.6.2 Immigrants and Regional Productivity Growth

Recall that aggregate productivity in region r is given by $A_{rt} = Z_{rt}^A (Q_{rt})^{\frac{1}{\sigma-1}}$ (see (2.17)). Hence, regional productivity growth is given by

$$\ln A_{rt} - \ln A_{rt-1} = \frac{1}{\sigma - 1} (\ln Q_{rt} - \ln Q_{rt-1}) + \ln Z_{rt}^A - \ln Z_{rt-1}^A$$

Using (2.23) we get that

$$\ln Q_{rt+1} = \frac{\iota + \lambda - 1}{\iota} \ln Q_{rt} + \ln H_{rt}^R + \ln \zeta_r + \ln Z_t^I.$$

Hence,

$$\ln A_{rt} = \frac{\lambda - 1 + \iota}{\iota} \ln A_{rt-1} + \frac{1}{\sigma - 1} \ln H^R_{rt-1} + \frac{1}{\sigma - 1} \ln \zeta_r + \frac{1}{\sigma - 1} \ln Z^I_{t-1} + \ln Z^A_{rt} - \frac{\lambda - 1 + \iota}{\iota} \ln Z^A_{rt-1}.$$

This suggests a specification of

$$\ln A_{rts} = \delta_s + \delta_t + \rho \ln A_{rt-1s} + \beta \ln H^R_{rt-1} + X'_{rt}\gamma + u_{rst},$$

where δ_s denotes a set of sector fixed effects to control for sector heterogeneity (which is not present in our theory), $X'_{rt}\gamma$ denotes a set of sectoral controls, which controls for regional differences in research technology ζ_r and the time fixed effect δ_t controls for the aggregate state of research efficiency $\ln Z^I_{t-1}$.

Changes in the supply of immigrants affect the provision of research human capital in region r, H^R_{rt}. Our model implies that (see (2.16))

$$\begin{aligned} \ln H^R_{rt} &= \varsigma + \ln L_r + \ln \left[\left(1 - \varpi_{rIt}\right) \left(h^R_N\right)^{\frac{1}{\theta}} \left(s^R_{rNt}\right)^{\frac{\theta-1}{\theta}} + \varpi_{rIt} \left(h^R_I\right)^{\frac{1}{\theta}} \left(s^R_{rIt}\right)^{\frac{\theta-1}{\theta}} \right] \\ &= \varsigma + \ln L_r + \frac{\theta - 1}{\theta} \ln s_{rNt} + \frac{1}{\theta} \ln h^R_N + \ln \left[\left(1 - \varpi_{rIt}\right) + \varpi_{rIt} \left(\frac{h^R_I}{h^R_N} \right)^{\frac{1}{\theta}} \left(\frac{s^R_{rIt}}{s^R_{rNt}} \right)^{\frac{\theta-1}{\theta}} \right] \end{aligned} \quad (2.30)$$

where ς is an inconsequential constant.[10] Log-linearizing the last term around the case of "no immigrants", i.e. $\varpi_{rIt} = 0$, we get that

$$\ln A_{rts} = \delta_s + \delta_t + \rho \ln A_{rt-1s} + \beta \ln L_{rt-1} + \beta \frac{\theta - 1}{\theta} \ln s^R_{rNt-1} + \frac{1}{\theta} \ln h^R_N + \alpha_{rt-1} \varpi_{rIt-1} + X'_{rt}\gamma + u_{rst}, \quad (2.31)$$

where

$$\alpha_{rt-1} = \beta \left(\left(\frac{h^R_I}{h^R_N} \right)^{\frac{1}{\theta}} \left(\frac{s^R_{rIt-1}}{s^R_{rNt-1}} \right)^{\frac{\theta-1}{\theta}} - 1 \right). \quad (2.32)$$

Equations (2.31) and (2.32) show that the relationship between productivity growth $\ln A_{rts} - \ln A_{rt-1s}$ and the share of immigrants ϖ_{rI}, depends crucially on the extent of comparative advantage. If immigrants have a comparative advantage in the research sector (i.e. $h^R_I > h^R_N$) and have a high employment shares in research $s^R_{rIt} > s^R_{rNt}$, our theory predicts a positive

[10] In fact, $\varsigma = \ln\left(\Gamma\left(1 - \frac{1}{\theta}\right)\right)$.

	(1)	(2)	(3)	(4)	(5)	(6)
	OLS	OLS	OLS	OLS	IV	IV
Immigrant share	0.566***	0.429***	0.450***	0.490***	0.742***	0.683***
	(0.068)	(0.105)	(0.110)	(0.162)	(0.129)	(0.186)
ln Native research share		0.156***	0.168***	0.169***		0.133***
		(0.036)	(0.038)	(0.038)		(0.044)
ln Pop		-0.118***	-0.116***	-0.115***		-0.114***
		(0.017)	(0.017)	(0.017)		(0.017)
High pre-immig skills				-0.012		
				(0.034)		
Immigrant share x High pre-immig skills				-0.074		
				(0.186)		
ln Productivity	0.324***	0.388***	0.388***	0.389***	0.399***	0.385***
	(0.019)	(0.025)	(0.025)	(0.025)	(0.023)	(0.026)
Urban share			-0.121	-0.115		-0.130
			(0.139)	(0.139)		(0.140)
Manufacturing share			-0.120	-0.108		-0.154
			(0.127)	(0.128)		(0.136)
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	4625	2162	2162	2162	2540	2115
R^2	0.445	0.539	0.539	0.539	0.520	0.537

Notes: Robust standard errors in parentheses.

Table 22: Immigrants and Productivity Growth (Equation (2.31))

relationship between the share of immigrants and productivity growth, i.e. $\alpha_{rt-1} > 0$. If immigrants and natives are identical in terms of their skills, productivity growth and the share of immigrants should not be related once the size of workforce L_{rt-1} and the research employment share of natives s_{rNt} (which under homogeneity will coincide with the employment share of immigrants) is controlled for.

In Table 22 we report the results of estimating (2.31), when we restrict α_{rt-1} to be constant across time and space. In column 1, we report the relationship between future productivity and the share of immigrants, when we only control for the current level of productivity. There is a strong positive relationship. Quantitatively, a change in the immigrant share by 0.1 percentage points, i.e. from say 0.1 to 0.2, increases regional productivity ten years later by 5%. In column 2, we estimate the specification in (2.31) without any regional controls. This specification is consistent with the model if there is no systematic heterogeneity in spatial research productivity ζ_r. The coefficient on the research employment share of

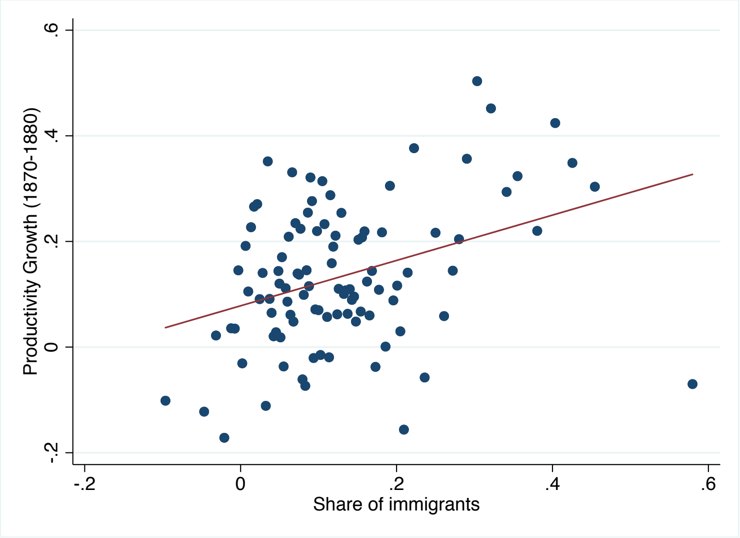


Figure 26: Immigrants and Productivity Growth

natives, $\ln s_{rNt-1}$, is positive and significant. In terms of structural parameters of the model, the coefficient should be equal to $\frac{1}{\sigma-1}\frac{\theta-1}{\theta}$. Empirically, total population size has a negative effect, even though the theory predicts a positive relationship. Below, when we consider patent activity instead of local productivity, we indeed find a positive relationship with local population. The relationship between future productivity and the immigrant share is slightly smaller compared. In column 3 we control for additional local characteristics, which we think of proxies for local research productivity. We utilize the share of the urban population and the manufacturing employment share. Both of these seems to be unrelated to future productivity growth and hence leave the coefficients on the other explanatory variables unchanged.

In the last two columns, we report the IV specification, where we instrument the actual immigrant share with the predicted share of immigrants as explained in Section 2.6.1 above. Using the instrument increases the coefficients slightly. As in the OLS specification, the coefficient on the immigrant share decreases slightly once we control for the share of researchers in the native population. Finally, in Figure 26, we depict the specification in column 3 of Table 22 graphically. We report a binscatter plot for 100 quantiles of the distribution of immigrant shares after all explanatory variables in columns 3 of Table 22 are controlled for. For ease of readability, we report productivity growth as the dependent variable. Figure 26 shows that the correlation between immigrants and productivity growth is robustly positive and not driven by particular outliers.

2.6.3 Immigrants and Regional Patent Activity

Our theory also makes predictions on the relationship between regional immigration inflows and the creation of patents. We adopt the following measurement approach. Let $\mathcal{P}_{rt}$ be the *stock* of patents filed in region r up to year t. We assume that Q_{rt} is proportional to the number of parents, i.e.[11]

$$Q_{rt} = \kappa \mathcal{P}_{rt}.$$

Hence, the model implies that (see (2.23))

$$\frac{\mathcal{P}_{rt}}{\mathcal{P}_{rt-1}} = \frac{Q_{rt}}{Q_{rt-1}} = \frac{Q_{rt-1}^{\iota+\lambda-1} H_{rt-1}^{R} \zeta_r Z_{t-1}^{I}}{Q_{rt-1}} = Q_{rt-1}^{\iota+\lambda-2} H_{rt-1}^{R} \zeta_r Z_{t-1}^{I}.$$

This implies that

$$\ln\left(\frac{\mathcal{P}_{rt}}{\mathcal{P}_{rt-1}}\right) = (\iota+\lambda-2)\ln Q_{rt-1} + \ln H_{rt-1}^{R} + \ln \zeta_r + \ln Z_{t-1}^{I}. \tag{2.33}$$

Using again the approximation for $\ln H_{rt-1}^{R}$ (see (2.30)), equation (2.33) suggests the regression

$$\ln \mathcal{P}_{rst} = \delta_s + \delta_t + \rho \ln \mathcal{P}_{rst-1} + \beta \ln L_r + \eta \ln s_{rNt} + \mu \varpi_{rIt-1} + X_{rt}'\gamma + u_{rst},$$

where the theory implies that $\rho = \iota + \lambda - 1$, $\beta = 1$, $\eta = \frac{\theta-1}{\theta}$ and $\mu > 0$ if and only if $\left(\frac{h_I^R}{h_N^R}\right)^{\frac{1}{\theta}} \left(\frac{s_{rIt-1}^R}{s_{rNt-1}^R}\right)^{\frac{\theta-1}{\theta}} > 1$. The time fixed effects control for the state of the research technology, $\ln Z_{t-1}^{I}$, and $X_{rt}'\gamma$ contains a set of observable regional characteristics which control for the systematic variation in research efficiency across space, ζ_r.

We report the empirical results in Table 23, which has the exact same structure as Table 22 above. Columns 1 - 3 report the OLS specification, where we only control for the

[11] The *flow* of new patents in year t, i.e. the number of patents filed at year t, N_{rt}^{Pat}, is therefore given by $N_{rt}^{Pat} = \mathcal{P}_{rt} - \mathcal{P}_{rt-1}$.

	(1) OLS	(2) OLS	(3) OLS	(4) OLS	(5) IV	(6) IV
Immigrant share	0.813***	0.144	0.182	-0.019	0.465**	-0.212
	(0.103)	(0.111)	(0.113)	(0.143)	(0.215)	(0.195)
ln Native research share		0.577***	0.600***	0.599***		0.641***
		(0.067)	(0.069)	(0.068)		(0.070)
ln Pop		0.188***	0.189***	0.185***		0.195***
		(0.057)	(0.057)	(0.056)		(0.058)
High pre-immig skills				-0.016		
				(0.036)		
Immigrant share x High pre-immig skills				0.368*		
				(0.197)		
ln Stock of patents	0.914***	0.755***	0.755***	0.754***	0.909***	0.756***
	(0.008)	(0.035)	(0.035)	(0.035)	(0.012)	(0.036)
Urban share			0.048	0.019		0.059
			(0.084)	(0.087)		(0.084)
Manufacturing share			-0.262*	-0.290**		-0.166
			(0.140)	(0.140)		(0.144)
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes
Observations	2641	1518	1518	1518	1732	1490
R^2	0.880	0.927	0.928	0.928	0.912	0.929

Notes: Robust standard errors in parentheses.

Table 23: Immigrants and Patent Activity

level of the patent stock $\mathcal{P}_{rst-1}$ (column 1), additional determinants of the supply of human capital (column 2) and additional regional characteristics (column 3). The difference between Table 23 and 22 is informative about the economic mechanism, in particular the extent of knowledge diffusion. While immigrants seems to have a robust positive effect of productivity growth, there is little evidence on a direct effect on patenting once the research share of the native population is controlled for. Hence, immigrants might have been less important for the *direct* increase in patent activity but affected productivity growth *indirectly* through knowledge diffusion to the native population who subsequently sorted into research-intensive occupations. The last two columns contain the IV specification, which are again qualitatively similar to the results using OLS. In Figure 27, we depict the specification in column 3 of Table 23 graphically.

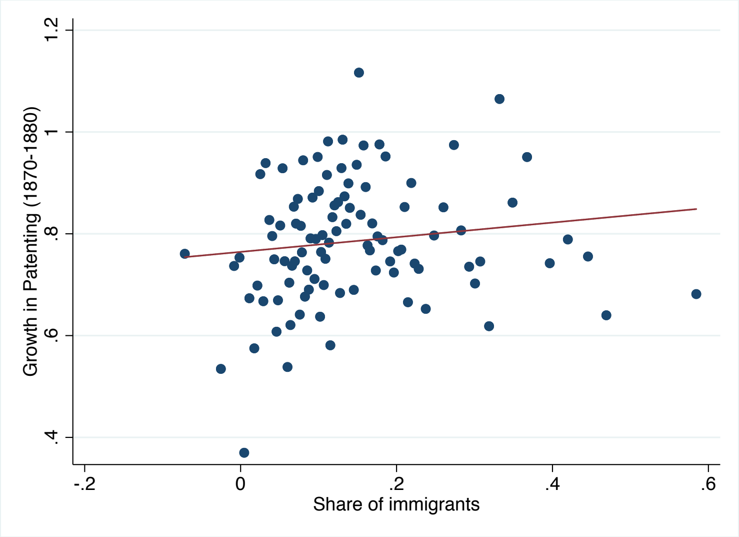


Figure 27: Immigrants and Patent Activity

2.6.4 Immigrants and Patent Novelty

We finally report the same regression but using spatial patent novelty as our left-hand-side variable. Our textual analysis provides a measure of *spatial idea novelty* for each region that we relate it systematically to the spatial inflow of immigrants. Columns 1 - 6 report the OLS specification, where we only control for the level of the patent novelty (column 1), additional controls (2-4) and the supply of human capital (column 6) and the IV specification with the level of patent novelty and the full set of controls and supply of human capital proxies. Almost all the specificationz yield a statistically significant role for the immigrants for local patent novelty that go beyond the scale effects and remain significant even after controlling for various proxies of human capital.

2.7 Counterfactuals

To conduct the counterfactuals we build on the procedure of Dekle et al. (2007) on conditioning the unobservables on actual data. We define $\hat{x} = x'/x$ and apply this definition to all the equilibrium equations of the model. For the trade module, equation 2.24,

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	OLS						IV	
Immigrant share	0.012***	-0.001	0.009***	0.010***	0.005***	0.003*	0.027***	0.006**
	(0.002)	(0.002)	(0.001)	(0.001)	(0.001)	(0.002)	(0.003)	(0.003)
ln Native research share		0.007***						0.005***
		(0.001)						(0.001)
ln Pop		0.003***	0.003***	0.003***	0.003***	0.003***		0.003***
		(0.000)	(0.000)	(0.000)	(0.000)	(0.000)		(0.000)
High pre-immig skills						-0.001*		
						(0.000)		
Immigrant share x High skills						0.004**		
						(0.002)		
Patent novelty	0.144***	0.111***	0.127***	0.214***	0.208***	0.207***	0.131***	0.113***
	(0.034)	(0.043)	(0.033)	(0.023)	(0.022)	(0.022)	(0.034)	(0.042)
Urban share					-0.001	-0.001		-0.002
					(0.000)	(0.000)		(0.002)
Manufacturing share					0.012***	0.012***		0.009***
					(0.001)	(0.001)		(0.003)
Industry FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	3282	1635	3282	6321	6321	6321	2113	1606
R^2	0.729	0.209	0.744	0.733	0.737	0.737	0.046	0.208

Notes: Robust standard errors in parentheses.

Table 24: Immigrants and Patent Novelty

$$\hat{w}^P_{rt}\hat{H}^P_{rt}w^P_{rt}H^P_{rt} = \sum_j \hat{x}_{ijt}x_{ijt}\hat{w}^P_{jt}\hat{H}^P_{jt}w^P_{jt}H^P_{jt}, \tag{2.34}$$

where

$$\hat{x}_{ijt} = \frac{\left(\frac{\hat{w}^P_r}{\hat{A}_{rt}}\hat{\tau}_{rj}\right)^{1-\varepsilon}}{\sum_r x_{ijt}\left(\frac{\hat{w}^P_r}{\hat{A}_{rt}}\hat{\tau}_{rj}\right)^{1-\varepsilon}}$$

and

$$\hat{A}_{rt} = \hat{Z}^A_{rt}\left(\hat{Q}_{rt}\right)^{\frac{1}{\sigma-1}}.$$

Now the innovation module using equations *(2.10) and (2.11)*

$$\hat{v}_{rt}v_{rt} = \frac{1}{\sigma}\frac{E_{rt}}{Q_{rt}}\frac{\hat{E}_{rt}}{\hat{Q}_{rt}} + \frac{\iota-1}{\iota}\left(\frac{\hat{v}_{rt+1}}{1+r_t}v_{rt+1}\right)^{\frac{\iota}{\iota-1}} \left.\frac{\hat{Q}^\lambda_{rt}\zeta_r\hat{Z}^I_t}{\hat{w}^R_{rt}}\frac{Q^\lambda_{rt}Z^I_t}{w^R_{rt}}\right)^{\frac{1}{\iota-1}} \iff$$

$$\hat{v}_{rt}v_{rt} = \frac{1}{\sigma-1}\frac{w^P_{rt}H^P_{rt}}{Q_{rt}}\frac{\hat{w}^P_{rt}\hat{H}^P_{rt}}{\hat{Q}_{rt}} + \frac{\iota-1}{\iota}\left(\frac{\hat{v}_{rt+1}}{1+r_t}\right)^{\frac{\iota}{\iota-1}} \left.\frac{\hat{Q}^\lambda_{rt}\hat{Z}^I_t}{\hat{w}^R_{rt}}\right)^{\frac{1}{\iota-1}} (v_{rt+1})^{\frac{\iota}{\iota-1}}\left(\frac{Q^\lambda_{rt}\zeta_r Z^I_t}{w^R_{rt}}\right)^{\frac{1}{\iota-1}}$$

Notice that, using equation 2.22 we have

$$H_{rt}^{R}=\frac{1}{\iota}\left(\frac{v_{rt+1}}{1+r_{t}}\frac{1}{w_{rt}^{R}}\right)^{\frac{\iota}{\iota-1}}\left(Q_{rt}^{\lambda}\zeta_{r}Z_{t}^{I}\right)^{\frac{1}{\iota-1}}Q_{rt}$$

so that

$$\hat{v}_{rt}v_{rt}=\frac{1}{\sigma-1}\frac{w_{rt}^{P}H_{rt}^{P}}{Q_{rt}}\frac{\hat{w}_{rt}^{P}\hat{H}_{rt}^{P}}{\hat{Q}_{rt}}+(\iota-1)\left(\hat{v}_{rt+1}\right)^{\frac{\iota}{\iota-1}}\quad\left.\frac{\hat{Q}_{rt}^{\lambda}\hat{Z}_{t}^{I}}{\hat{w}_{rt}^{R}}\right)^{\frac{1}{\iota-1}}\frac{H_{rt}^{R}w_{rt}^{R}}{Q_{rt}}\Longleftrightarrow$$

$$\hat{v}_{rt}v_{rt}Q_{rt}=\frac{1}{\sigma-1}w_{rt}^{P}H_{rt}^{P}\frac{\hat{w}_{rt}^{P}\hat{H}_{rt}^{P}}{\hat{Q}_{rt}}+(\iota-1)\left(\hat{v}_{rt+1}\right)^{\frac{\iota}{\iota-1}}\quad\left.\frac{\hat{Q}_{rt}^{\lambda}\hat{Z}_{t}^{I}}{\hat{w}_{rt}^{R}}\right)^{\frac{1}{\iota-1}}H_{rt}^{R}w_{rt}^{R}\tag{2.35}$$

Same equations imply that

$$v_{rt}Q_{rt}=\frac{1}{\sigma-1}w_{rt}^{P}H_{rt}^{P}+\frac{\iota-1}{\iota}H_{rt}^{R}w_{rt}^{R}$$

With this equation we can calibrate the initial

$$v_{r0}Q_{r0}$$

with knowledge of $w_{r0}^{P}H_{r0}^{P},H_{r0}^{R}w_{r0}^{R}$.

Also notice that equation 2.22 in changes is

$$\hat{H}_{rt}^{R}=\left(\frac{\hat{v}_{rt+1}}{\hat{1+r_{t}}}\frac{1}{\hat{w}_{rt}^{R}}\right)^{\frac{\iota}{\iota-1}}\left(\hat{Q}_{rt}^{\lambda}\hat{Z}_{t}^{I}\right)^{\frac{1}{\iota-1}}\hat{Q}_{rt}\tag{2.36}$$

The second equation for the innovation module (2.23) can be directly written in changes

$$\frac{\hat{Q}_{rt+1}}{\hat{Q}_{rt}}=\hat{i}_{t}=\left(\hat{H}_{rt}^{R}\hat{Z}_{t}^{I}\hat{Q}_{rt}^{-(1-\lambda)}\right)^{1/\iota}\tag{2.37}$$

Now we can do the full DEK! Given $v_{r0}Q_{r0}$ that we obtain from some data we can solve for $\hat{Q}_{rt+1},\hat{v}_{rt+1},\hat{w}_{rt}^{R},\hat{w}_{rt}^{P},\hat{H}_{rt}^{R},\hat{H}_{rt}^{P}$ using the above equations in changes and the two labor supply equations Conditional on the initial levels of $\left\{w_{rt}^{j}H_{rt}^{j}\right.$ for $j=P,R$ and $t=0$ and changes in $\left\{\hat{Z}_{rt}^{A},\hat{Z}_{t}^{I}\right\}$ the changes in $\left\{\hat{Q}_{rt},\hat{v}_{rt},\hat{w}_{rt}^{R},\hat{w}_{rt}^{P},\hat{H}_{rt}^{R},\hat{H}_{rt}^{P}\right\}$ can be determined with the solution of equations 2.34, 2.35, 2.37, and 2.36, and the labor supply equations for $\left\{\hat{H}_{rt}^{R},\hat{H}_{rt}^{P}\right\}$.

2.8 Conclusions

We have developed an empirical and theoretical framework to analyze the role of human capital and spatial policies on economic growth. Our big data historical approach allows us to analyze decades of data on the American economy and the associated effects of the influx of immigrants. We couple these data with a model of forward looking innovating firms that allows us to evaluate the empirical data using structural relationships that arise from the theory. In future work we plan to exploit the micro aspect of the data to fully understand the process of knowledge creation by immigrants and to provide more definitive conclusions on the impact of immigration on American growth.

Chapter 3

American Intergenerational Mobility in History and Space

Rodrigo Adao, Costas Arkolakis, Sun Kyoung Lee[1]

[1] We thank Jack Liang for truly outstanding research assistance. All remaining errors are our own.

American Intergenerational Mobility in History and Space

Rodrigo Adao — Chicago Booth

Costas Arkolakis — Yale University

Sun Kyoung Lee — Columbia University

Abstract

We exploit advances in data digitization and machine learning to study intergenerational mobility in the United States before World War II. Using machine learning techniques we construct a massive database for multiple generations of fathers and sons. This allows us to identify “land of opportunities": locations and times in American history where kids had chances to move up in the income ladder. Our massive sample allows us to extrapolate income variation from occupations and demographic and geographic characteristics. We find that intergenerational mobility elasticity relatively stable during 1880-1940; there are regional disparities in terms of giving kids opportunities to move up, and the geographic disparities of intergenerational mobility have evolved over time. Our findings and descriptive analyses do not identify the causal mechanisms of intergenerational mobility, but in future work, we plan to evaluate the impact of policies in explaining these patterns.

3.1 Introduction

Economic history is in the midst of a “data deluge”. The advancement of large-scale data collection and digitization techniques coupled with the increasing availability of massive administrative datasets enables researchers to reexamine economic history with a new quantitative approach. We exploit the new data availability and modern artificial intelligence-based record linking techniques to document intergenerational mobility in the United States. While extensive literature investigates whether the United States is true “land of opportunity” (e.g. Olivetti and Paserman (2015), Becker and Tomes (1986), Mazumder (2005), Solon (1992)), lack of limitations in data and scarcity of data has left the results somewhat debatable.

In this paper, we revisit this question and examine the intergenerational mobility in the United States in the late nineteenth century and early twentieth century. Relative to the existing literature, we exploit a machine-learning approach to link complete-count US federal demographic censuses across time to track the same individuals over time for almost one hundred years. Our approach has two main advantages relative to existing literature in historical intergenerational mobility. First, instead of using samples of individuals, we use the complete-count US population census with socioeconomic information such as occupation. Second, we develop comprehensive record linking algorithms and methodologies that improve the quality of matches. Third, we study multiple periods of intergenerational mobility spanning for almost one hundred years and give a complete account of the evolution of the social ladder in the US.

We document a small improvement in intergenerational mobility in the United States between 1880 to 1940. The improvement, is mostly reflected in smaller probability that childen of wealthy parents will be in the top quartile 20 or 30 years after and by a higher corresponding probability that childen of fathers around the median incomes would be in the top income quartiles.

Linking across historical US census records poses three main challenges; first, there are no time-invariant individual identifiers (e.g. social security number); second, lack of time-invariant individual identifiers means linking records across different sources is computationally extremely intensive and manual hand-linking within consistent record linking rules by humans is almost infeasible; third, the federal census did not ask income information until the Year 1940 that income information for pre-1940 record is not available.

We tackle the first and second challenges by introducing a machine-learning based automated record linking approach. Firstly, given the absence of time-invariant individual identifiers, we combine parental information and individual-level information to link individual records across years. This involves a careful design and implementation of machine-learning based automated record linking using time-invariant information such as first and last names (women typically change their last names after marriage, and details regarding women's record linking are available in Appendix), age and birthplaces and household-level information such as parental birthplaces, ages, and given names. Our contribution is to develop comprehensive record linking algorithms and methodologies that enables researchers to link not only individuals, but also multi-generations of individuals (e.g. fathers and daughters, fathers and sons, grandfathers-fathers and sons), so that we constructa truly longitudinal database of individuals of both gender and across time, and across generations.

We approach the third challenge of income information unavailability for pre-1940 US demographic censuses by extrapolating income profiles using rich individual characteristics and income information in later US censuses. Unfortunately, income collection in the census has only started from 1940 for wage income and therefore historical microdata containing earnings, along with individual-characteristics such as names and ages are extremely rare. As we lack comprehensive income information for historical periods (i.e. pre-1940), we use 1940 and 1950 income information from US demographic census to construct predicted measures of position in the income distribution based on individual demographic characteristics. 3.2.3 discusses income extrapolation related details and step-by-step implementation.

Our paper expands the existing literature by linking parents and children across all years from 1850 to 1940 and creates intergenerational longitudinal datasets of fathers and sons for almost one hundred years. Our paper is closely related to the work of Olivetti and Paserman (2015), who estimate historical intergenerational elasticities between fathers and children by creating *pseudo*-links across generations using 1 percent extracts from the decennial censuses of the US between 1850 and 1940. Relative to Olivetti and Paserman (2015), we create a true panel of fathers and sons using complete-count decennial censuses of the US of the same period. Feigenbaum (2015) studies whether severe economic downturns affect intergenerational economic mobility by estimating rates of intergenerational mobility during the Great Depression in American cities with varying degree of economic downturns. Similar to our effort, Feigenbaum (2015) also develops a matching algorithm to link parents and children across censuses and link 1920 and 1940 US demographic censuses. He identifies differential directed migration as a key mechanism of explaining lowered intergeneration mobility for sons growing up in severely economic downturn-hit cities.

The rest of the paper proceeds as follows. 3.2 describes the data construction methodology and discusses measures of income imputation. 3.3 presents the main results and 3.5 concludes.

3.2 Record Linking Methodology and Imputation of Incomes

In this section, we describe record linking procedure and income extrapolation in detail used in our analysis.

3.2.1 Record Matching Overview

The current census matching methods widely used in economic history can be classified into three categories. The following is the (non-exhaustive) list the method along with papers that adopt the corresponding record linking techniques:

- Traditional iterative decision tree method: Abramitzky et al. (2012b), Nix and Qian (2015)
- Automated Probabilistic Algorithm: Abramitzky et al. (2018)
- Supervised discriminative learning method: Goeken et al. (2011), Feigenbaum (2015)

3.2.1.1 Traditional Iterative Decision Tree Method

As discussed above, the first approach uses a traditional method of record linking and it was widely implemented by Ferrie (1996), Long and Ferrie (2013), and Abramitzky et al. (2012b). This method adopts an iterative matching technique and is largely drawn from works on data-mining techniques. This iterative matching approach relies on probabilistic record linking algorithms with an additive point system that assesses the similarity of individual records and household information. Researchers specify a set of rules defining what is a true match in practice. For example, in Abramitzky et al. (2012b), records with the same standardized names, with the same birthplace and their age gap is within 3 years are considered as a true match. The decision tree method is simple and easy to implement.

The critique of this approach often involves the eliminations of duplicates and linkage procedure itself. For example, suppose there are multiple potential links—the first potential link has exactly the same first name and last name strings, whereas the age is off one year, and the second potential link has slightly different name spellings or diminutive of first names and the age is the exact age match. The reiterative approach would reject all potential links with ambiguity. This elimination of all potential matches not only lowers the overall match rate, but it may also introduce a systematic bias in linked data (for example, people with relatively uncommon combinations of matching criteria such as first and last names may be overrepresented in the matched data and vice versa).

This method's limitation of became motivation for further record linking method development. While traditional iterative decision tree method does not assign "appropriate" weight differences in name spelling and age when comparing two records, the methods that I discuss in the following combines these differences (e.g. differences in names, and differences in age/birth year, differences in family characteristics) in linking records.

3.2.1.2 Automated Probabilistic (EM) Algorithm

Abramitzky et al. (2018) uses the Expectation-Maximization (EM, henceforth) algorithm to compute the probability that each two records correspond to the same individual. The EM algorithm is a standard technique in the statistical literature that weighs name and age difference for record linkage. The primary difference between EM algorithm relative to iterative method is the weight treatment of discrepancies such as name spelling differences and age differences.

3.2.1.3 Machine Learning Method (This Paper)

The machine-learning algorithm uses a "training dataset" (i.e. a sample hand-linked records with samples of "true" and "false" matches with associated characteristics such as age, first, middle, and last name, birthplace) to train an algorithm to link records given discrepancies in record features (e.g. age differences, birthplace differences, name differences). Through training data, researchers train the algorithm how to identify a potential match based on the patterns of "true" and "false" matches. 3.2.2 explains record linking procedure in details.

3.2.2 This Paper: A (Supervised Discriminative) Machine Learning Approach for Record Linking

We implement a supervised discriminative machine learning approach to link historical records. The essence of this approach is that researchers use training data (as "teaching-material") to train the algorithm on how to identify the potential match based on certain discrepancies in the data (for example, Heinrich Engelhard Steinweg, the founder of prominent piano manufacturing company, *Steinway & Sons*, anglicized his names into "Henry E.

Steinway." Therefore, in linking his records across censuses, string comparison measures called Jaro-Winkler distance of his first (Heinrich vs. Henry), middle (Engelhard vs E.) and last name (Steinweg vs Steinway) would show name discrepancies) even if his birth year and birthplace may be the same across different records).

We create a linked-individual longitudinal database across different census years. We exploit the complete transcription of decennial federal census records from 1850 to 1940 except for 1890 (which was lost due to fire). For each individual, the data include demographic and geographic information (sex, age, race, place of birth, place of residence), as well as labor market outcomes (occupation and industry). Similar efforts of linking records using machine learning methods have been made by Goeken et al. (2011) that built the IPUMS linked individual samples, Feigenbaum (2016) has undertaken the task of linking historical records of 1915 Iowa State Census to their adult-selves in the 1940 Federal Census.

Relative to the mentioned work, this project is far more extensive in the scope of matching as it involves complete-count census records from 1850 to 1940. We teach a machine to learn to predict whether two records look "true" links of the same individual or not based on a set of observable features. In particular, we implement a supervised learning algorithm where the presence of outcome variable ("true" or "false" links) guides the learning process; in other words, the end goal is to use the inputs to predict whether the potential links are "true" or "false" matches.

Between two census years, we first link (1) males between age 0 and 16 ("sons") with their father present in the same household (sharing the same household identification number) in earlier year census and (2) now adult-selves of sons whose age is 30-36 (for 20-year gap data) or 30-46 (for 30-year gap data) which we call "adult(selves)" in later year census. Once we link "sons" and "adult-selves" between two censuses, we then go back to the earlier census and identify "fathers" of linked "sons".

3.2.3 Methodology to Construct Position in the Income Distribution

Despite the richness of the linked census, individual income is not available for pre-1940 population census. To deal with the absence of income information for pre-1940 census, we use 1940 and 1950 income information to construct predicted measures of position in the income distribution based on individual demographic characteristics. Due to the creation and the division of counties as the United States expanded its territory we use the geographical definitions of 1940 State Economic Areas (SEA) provided by the Minnesota Population Center, which allow us to create consistent geographical boundaries throughout our sample. Our sample includes a maximum of 359 SEAs. We explain the implementation steps of our methodology below.

The projection of occupation on individual characteristics can be represented as

$$\ln y_i^0 = X_i \gamma + e_i, \tag{3.1}$$

where y_i^0 is income and X_i is a vector of attributes which includes the following variables: age dummies, race dummies (white, black, other), US born fixed effect, urban fixed effect, literate fixed effect (it exists in every year prior to 1940, but we need to use school attainment to construct it in 1940), occupation fixed effect, and finally an SEA fixed effects.[1] For the projection we pre-select in the 1940 census a sample of males, aged 16-60, that are employed and with available information on income, and county (ultimately aggregated to SEA) information.

For each individual in the censuses prior to 1940 we compute a predicted income given the individual's observed characteristics X_i and the estimated coefficients $\hat{\gamma}^{1940}$, $\ln \hat{y}_i = X_i \hat{\gamma}^{1940}$. For each income measure we then compute the position of the individual in the distribution of income. We repeat this for our sample of individuals in the years 1950, 1960, 1970 to cross-validate our methodology (1940, 1960, 1970 when we use the 1950 census). We are of

[1] We experimented with incorporating industry dummies but the predictive power was only slightly affected while the sample was reduced due to the unavailability of industry information for many individuals.

course aware of the issue of missing self-employed farmer incomes from the 1940 census that the literature has raised (see Olivetti and Paserman (2015)). In the Appendix we reconstruct all our statistics with the 1950 census that contains total income, and not only wages, and obtain very similar results as we discuss below.

The following table discusses the quality of income imputation using 1940 (wage income/ INCWAGE) and 1950 (total income/ INCTOT) income data as a basis. We report the R^2 of our prediction regression (in-sample) by using different explanatory variables in the following specifications (1) age, US born, literacy status, race, urban (for 1940 only), Occupations and characteristics as in specification (1), State and characteristics as in specification (2), and, finally, only for 1940, SEA and characteristics as in specification (3). As it is evident while individual characteristics have strong explanatory power, occupation has the largest predictive power in explaining (in-sample) income in both years. This investigation guides our baseline specification for the income extrapolation, which is richer specification (4) for 1940.

Table 31: Model Fit: Manufacturing Employment - Alternative Specifications

Specification	1940 R^2	1950 R^2
(1) Age, US born, literacy, race, urban (1940 only)	0.269	0.200
(2) Occupation and characteristics in (1)	0.424	0.395
(3) State and (2)	0.438	0.414
(4) SEA for 1940 and (2)	0.449	n/a

Using our baseline specification we next predict in-sample and out-of sample individual incomes in Table 32. We do this as means to compare the performance of our extrapolation methodology and the capacity of the 1940 (full) census versus the 1950 census (10% sample) to better predict income. We see that both samples allow us to extrapolate income, based on various characteristics, fairly well as those extrapolations allow for a good fit not only in-sample (i.e. for the data of that year, indicated with bold) but also out-of-sample by looking at how well we predict income in other years, typically with an R^2 upwards of 0.4

around the in-sample fit. Notice that due to the 70-year confidentiality restriction we do not have access to the names in the 1950 and onwards censuses and the samples which restricts the geographic information we can access and the sample is also considerably smaller. Since the fit is not very different but the underlying sample we base our extrapolation is two orders of magnitude larger for 1940 than 1950 and can contain characteristics like urban, and SEA geographic location fixed effects we use this as our prefer specification. In the Appendix we report the income distribution for each year based on the two extrapolations which are surprisingly similar, with a discernible difference for between the 20th and the 30th percentile (lead by possible difference in the predicted farmer income).

Table 32: Income Extrapolation: Predictive Power of Income Extrapolation based on 1940 and 1950 income in and out of sample.

Year	1940 R^2	1950 R^2	# of Sample
1940	**0.431**	0.412	26,058,449
1950	0.343	**0.395**	109,483
1960	0.447	0.466	2,031,270
1970	0.463	0.490	474,837

Notes: The table illustrates the predictive power of our extrapolation in- and out-of-sample for different samples. US-born white males with ages between 30-36 and a father of 16-40 years older than the son, with valid variables (occupation, age, urban, race, literacy, county in SEAs) linked to fathers by census. Income percentiles constructed from occupations based on income data in 1940 and 1950. See Section 3.2 for desciption.

We make two notes regarding our methodology. First, there is measurement error based on unobserved characteristics of individuals. The residual in (3.1) implies that we are missing a component of income that cannot be predicted with the vector X_i. The measurement error is classic by construction, but it is unlikely to be uncorrelated between father and son due to correlated unobservable characteristics. Second, the return to observable characteristics may change over time. We plan to address this by using better wage data by observable industry and occupation such as the data by Preston and Haines (1991) used by Olivetti and Paserman (2015).

Notice, finally, that the variation in the price index across regions could generate differences between the individual's position in the distribution of nominal income and real income. This is a problem only to the extent that this varies systematically with the position of the father in the income distribution. Chetty et al. (2014) report that adjusting for the price index does not affect their estimates.

3.2.4 Final Sample Selection and Summary Statistics

Based on the results above, we select a sample of individuals with specific information from the population census. We focus on white US-born males as other racial groups tend to be under-represented in the linked data and/or certain racial groups in historical periods account for relatively small proportion.[2] We also consider individuals of age 30-36 with valid occupation code and available demographics (age dummies, race dummies, US born dummy, urban dummy, literate dummy). Some individuals with "valid" occupations (i.e., non NA/missings) will be dropped as their occupation/industry does not exist in 1940. The next table illustrates details on the census sample size for adults of age 16-60 in our censuses. From a male population of tens of millions the census gives useful information for more than ten million individuals per year that contain all our relevant variables. Table 33 shows the sample in every year and the sample given conditional on the availability of different variables.

Table 33: Population Census Samples: Alternative Specifications

	Census Sample Size					
Variable	1880	1900	1910	1920	1930	1940
Male population	26,909,609	38,760,183	47,611,800	53,901,769	62,104,510	66,198,373
Aged 16-60	15,153,954	22,278,712	28,814,166	32,304,049	37,932,346	41,987,248
Valid occupation	14,064,328	20,144,159	26,811,926	29,634,769	34,368,103	36,056,074
Other valid variables	12,110,640	18,114,656	22,726,061	21,872,691	25,803,132	35,447,377

Out of these we restrict our sample to young working age men of age 30-36. We look for

[2] Abramitzky et al. (2019) notes "transcription differences between 1940 Federal Census are especially high for the foreign born from non-English speaking countries" which implies that the record linking procedure may induce US-born (whose name is more likely to be Anglicized) to be overrepresented in the linked data).

the same individuals in previous censuses using our machine learning techniques. We use four 20 year periods (1880-1900, 1900-1920, 1910-1930, 1920-1940) and three 30 year periods (1850-1880, 1880-1910, 1900-1930, 1910-1940).

Based on the sample of individuals linked in the initial year, we select sample of fathers with same valid variables that are 16-40 years older than their sons. Our methodology implies that we have a position on the income distribution of all males in the initial year (father) and in the final year (son). For each pair, we have the location (state and county) in the initial year when son and father are in the same household. We also have the place of residence of the son in the final year. That allow us to look at the geographic variation of intergenerational mobility in our sample, something that previous approaches where typically unable to do either due to small sample restrictions or due to lack of geographic identifiers. The next table presents summary statistics for the incidence of matches as a share of available data with age and demographics.

Table 34: Twenty Years Matched Sample Size: Alternative Specifications

Matched Sample Size				
Variable	1880-1900	1900-1920	1910-1930	1920-1940
30-36 age, white males, US born	2,630,624	3,953,411	4,779,392	5,795,011
Linked in initial year	0.334	0.219	0.391	0.233
Linked to a father	0.209	0.142	0.288	0.178
Valid occupation	0.196	0.133	0.276	0.166
Other valid variables	0.162	0.090	0.169	0.121

There is a systematic overestimate of the income of the lowest percentiles and underestimate for the largest, though the bias is, in general, small. In future work, we plan to use this information to improve our matches by enriching our training dataset with additional high quality matches.

3.3 Results

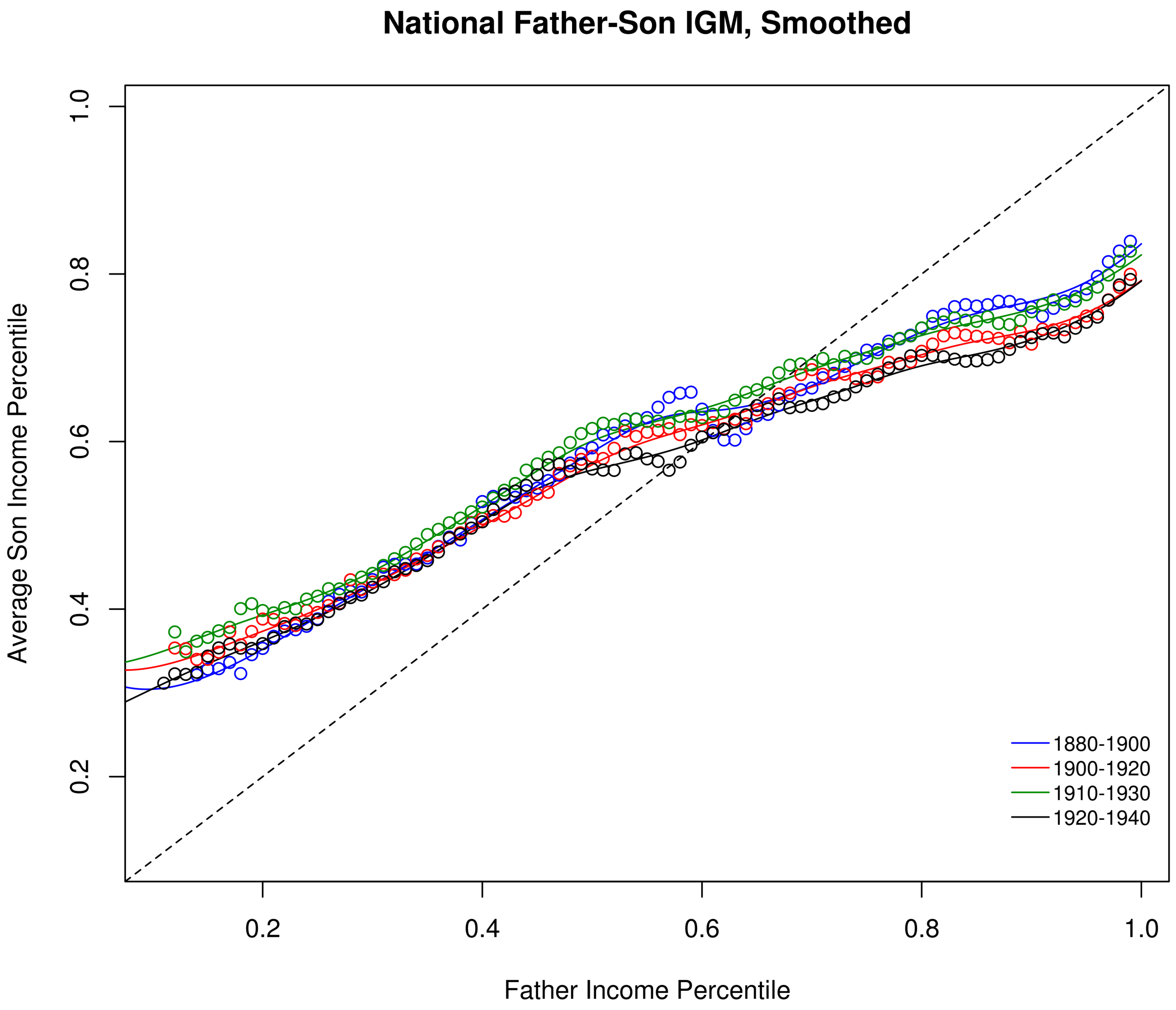
We now construct intergenerational mobility statistics in our sample for all our available matched years with 20 years matching intervals. We chose to present statistics by regressing the income percentile of sons on fathers. To create a smoothed measure we divide the sample of sons in 100 bins according to the position of their fathers in the income distribution in the initial year.

Among the individuals in each bin, we compute the following variables using the income percentile of sons in the final year: (i) average income position, $E\left[\hat{P}_i^t|\hat{P}_{i,father}^{t-20}=p\right]$. This is very similar to estimating directly the univariate regression of the percentile of the son on the percentile of the father; (ii) probability of being in the top quantile, $Pr\left[\hat{P}_i^t>.75|\hat{P}_{i,father}^{t-20}=p\right]$; (iii) probability of being in the bottom quantile, $Pr\left[\hat{P}_i^t<.25|\hat{P}_{i,father}^{t-20}=p\right]$; (iv) dispersion in income position, $StDev\left[\hat{P}_i^t|\hat{P}_{i,father}^{t-20}=p\right]$.[3]

Figure 31 illustrates the average son's income position on the father's percentile. The fact that the lines for each interval are crossing the 45 degrees line and are above the line in the lower percentiles and below that for the higher percentiles indicates mobility. For comparison, the case of no mobility is represented by the 45 degrees line that indicates that the expected son's percentile will be identical to the father's percentile. The graphs also show that intergenerational mobility has increased the first decades of the 20th century: sons of very rich fathers are less likely to remain rich. However, looking at average does not reveal the full picture of the dynamics of the American economy as it masks the distribution of these intergenerational movements. To capture those, we look at the probabilities of moving to the top and bottom quartiles in the next graph.

[3] In practice, to construct the points and the interpolation lines in our graphs we proceed as following. For the points: (i) For each linked pair, we multiply the father's income percentile by 100 and round below to floor to get an integer between 0 and 99. (ii) Divide by 100 to get a number in 0, ..., 0.99. (iii) this gives us N bins, typically around 95. (iv) For each bin, compute the average son's income pctile (v) plot the pair in the associated color for each year. For the lines: (i) on the full sample compute local polynomials. In particular, we use the function called locpoly in R using a Gaussian kernel and a bandwidth of 0.05 (2) Fit this local polynomial to a grid [0, 0.01, 0.02, ..., 0.99] (iii) Plot the fitted values of the polynomial in the associated color.

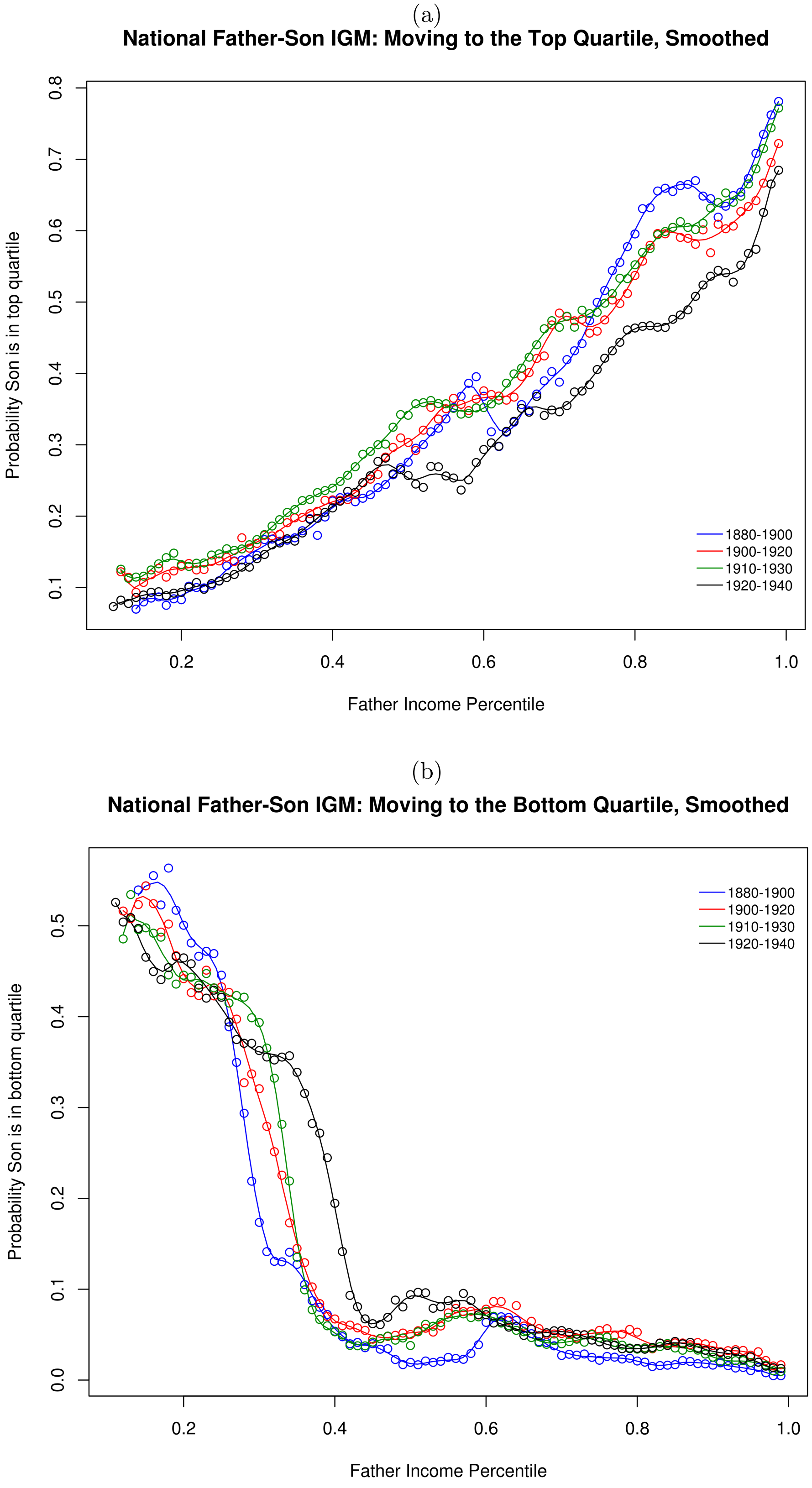
Figure 31: Intergenerational Mobility Coefficients: Son's Average Income Position



Notes: The figure plots average income position of father and son. US-born white males 30-36 and a father of 16-40 years older than the son, with valid variables (occupation, age, urban, race, literacy, county in SEAs) linked to fathers by census. Income percentiles constructed from occupations based on income data in 1940.

Figure 32 plots the probability that a son ends up in the top (left panel) and bottom (right panel) quartile of the income distribution in the final year of the linked sample. In these two graphs we see much more discernible differences. In particular, while the likelihood of ending up in the top quartile has remained roughly constant for the bottom half of the distribution, the kids with parents below the median income, the likelihood has decreased for kids of the most affluent parents going from about 70% around the 90th percentile in the 1880-1900 percentile to slightly above than 55% in the 1920-1940 linked census.

Figure 32: Intergenerational Mobility Coefficients: Son's Probability of Being in the Top and Bottomn Quartile.



(a)

(b)

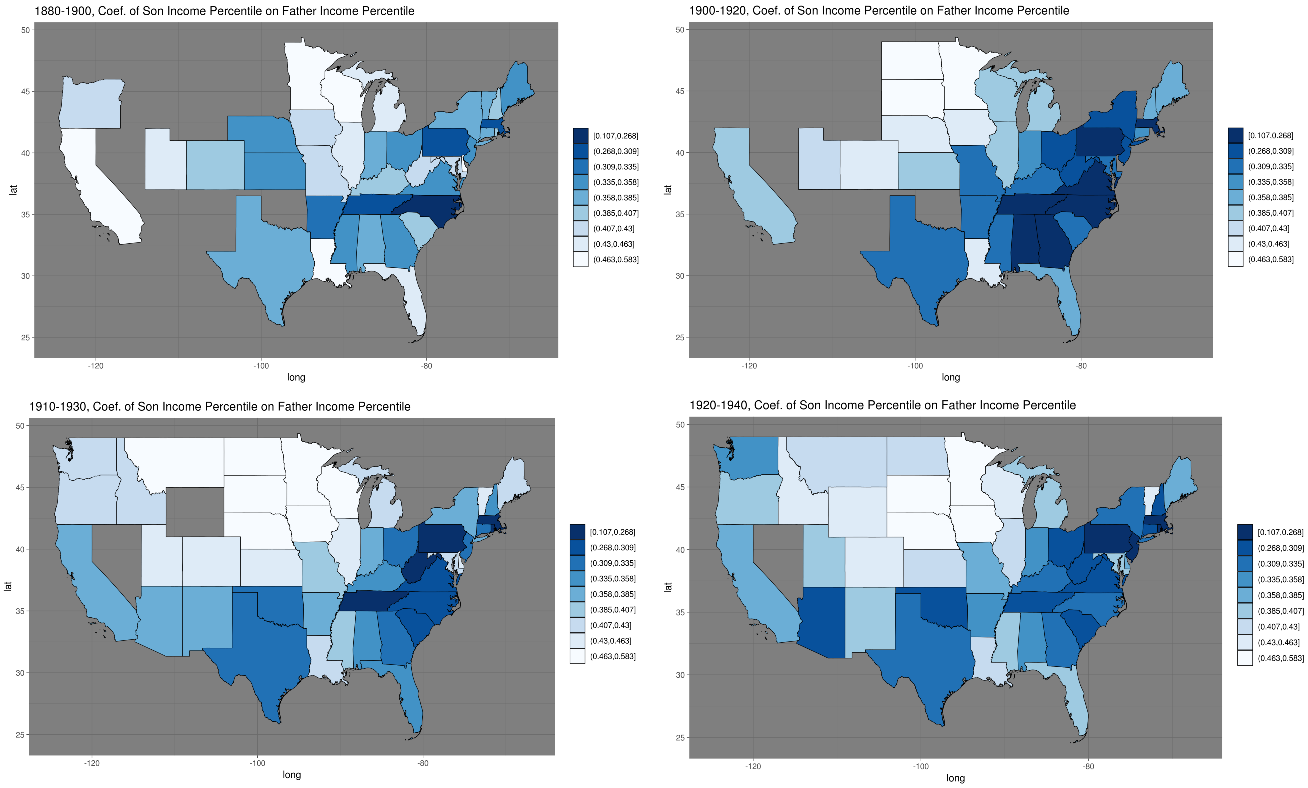
3.4 Historical Intergenerational Mobility in Space

Exploiting our large sample, we now study intergenerational mobility in space by looking at geographic variations across states and SEAs over 20-year matched samples of fathers and sons, where the location is determine by the location of the household of the father where the son was born. Figure 33 plots IGM coefficients across states for the four matched samples. A darker color indicates a lower IGM coefficient and, thus, more intergenerational mobility. We only plot states where we have more than 1000 observations within each state. That effectively implies that a large part of the thinly populated mid- and western states will be missing for the first matched samples.

There are at least two noticeable patterns that arise: one cross-sectional and one at the time series. First, the Northeast of the United States appears consistently as the area with the most income mobility with the midwest as the area with the lowest. Second, there is a substantial variation on the mobility over time. Mobility improves in many areas and especially in the west and the South-West regions. Largely, however, the US appears to be a place of very heterogeneous intergenerational mobility opportunities, in accordance with more recent findings by Chetty et al. (2014).

Figures 34 and 35 zoom in at the SEA level and present intergenerational mobility coefficient and the probability that a son is in the 4th income quartile conditional on the father being on the 1st quartile, respectively. We only plot SEAs where we have more than 100 observations within each SEA. The same patterns arise in these two figures. The Northeast appears as the champion of intergenerational mobility in the historical United States with some regions having very high intergenerational mobility coefficient and an impressive more than 50 percent change of rich kids with poor parents. However, even that early on in the American history the thinly populated Western United States still appear to develop an environment of high intergenerational mobility with very high expected incomeof sons with poor parents by 1940 in many SEAs.

Figure 33: Intergenerational Mobility Coefficients Across States: Son's Average Income Position

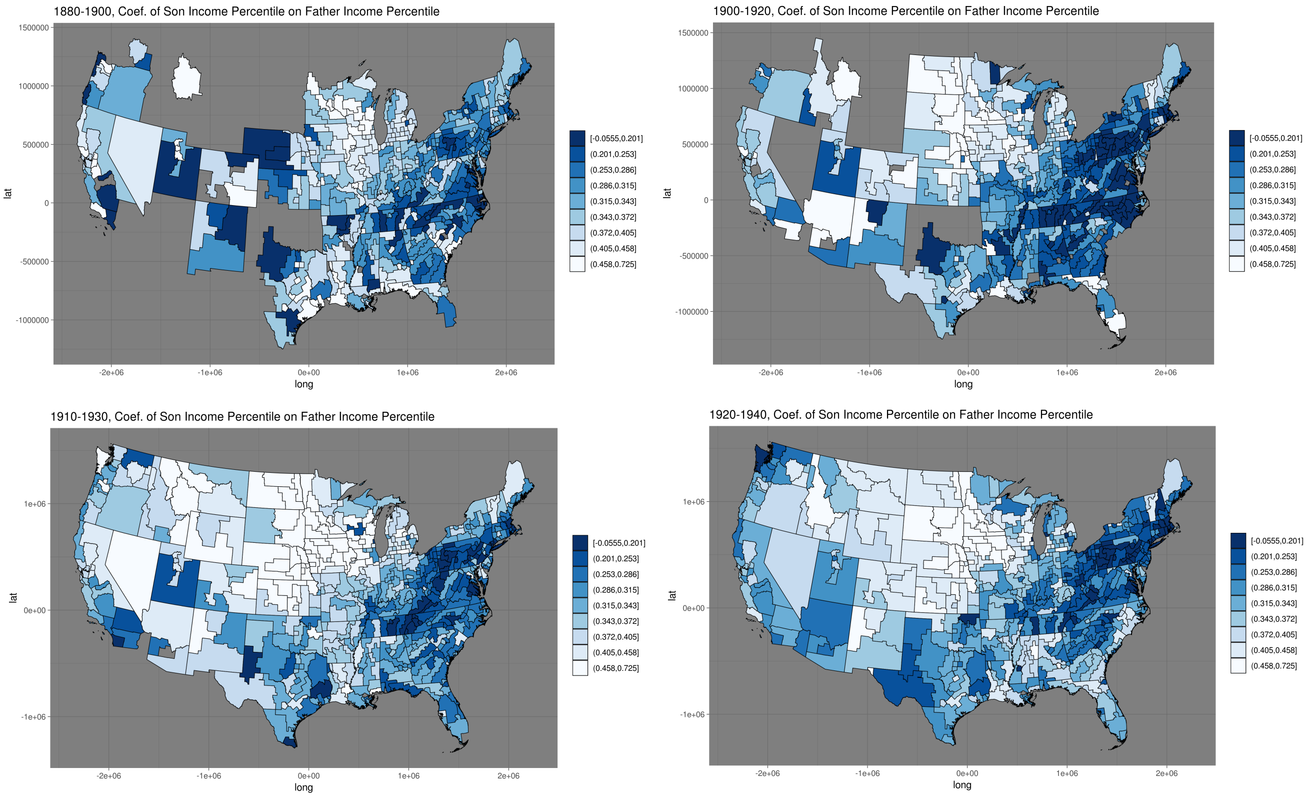

1880-1900, Coef. of Son Income Percentile on Father Income Percentile
1900-1920, Coef. of Son Income Percentile on Father Income Percentile
1910-1930, Coef. of Son Income Percentile on Father Income Percentile
1920-1940, Coef. of Son Income Percentile on Father Income Percentile
lat
long
50
45
40
35
30
25
-120
-100
-80
[0.107,0.268]
(0.268,0.309]
(0.309,0.335]
(0.335,0.358]
(0.358,0.385]
(0.385,0.407]
(0.407,0.43]
(0.43,0.463]
(0.463,0.583]

Notes: The figure plots average income position of father and son across SEAs. US-born white males 30-36 and a father of 16-40 years older than the son, with valid variables (occupation, age, urban, race, literacy, county in SEAs) linked to fathers by census. Income percentiles constructed from occupations based on income data in 1940.

3.5 Conclusion

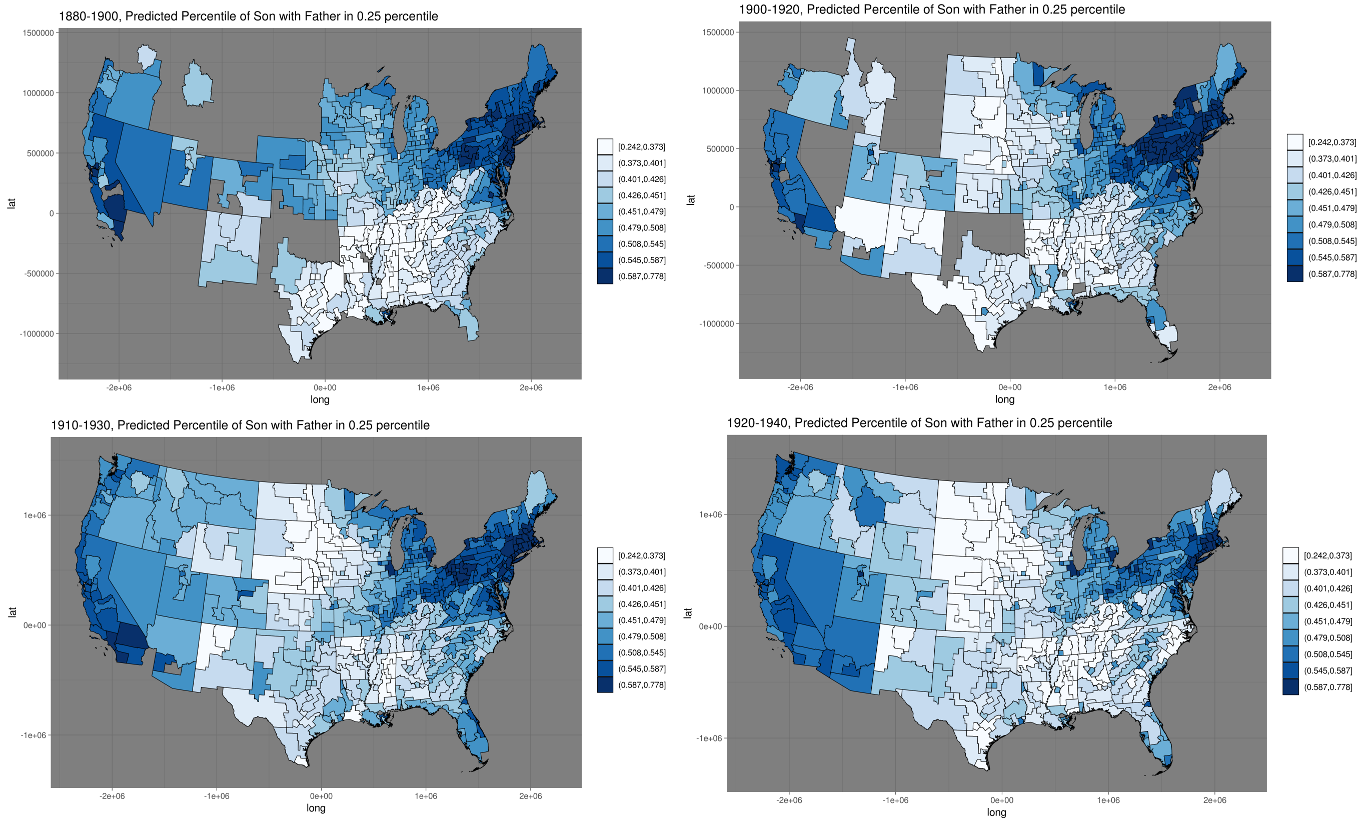
We find that intergenerational mobility elasticity relatively stable during 1880-1940. There are regional disparities in terms of giving kids opportunities to move up, and the geographic disparities of intergenerational mobility have evolved over time. Our findings and descriptive analyses do not identify the causal mechanisms of intergenerational mobility, but in future work, we plan to evaluate the impact of policies in explaining these patterns.

Figure 34: Intergenerational Mobility Coefficients Across SEAs: Son's Average Income Position



Notes: The figure plots average income position of father and son across SEAs. US-born white males 30-36 and a father of 16-40 years older than the son, with valid variables (occupation, age, urban, race, literacy, county in SEAs) linked to fathers by census. Income percentiles constructed from occupations based on income data in 1940.

Figure 35: Intergenerational Mobility Coefficients Across SEAs: Predicted Percentile of Son with Poor Father


1880-1900, Predicted Percentile of Son with Father in 0.25 percentile
1900-1920, Predicted Percentile of Son with Father in 0.25 percentile
1910-1930, Predicted Percentile of Son with Father in 0.25 percentile
1920-1940, Predicted Percentile of Son with Father in 0.25 percentile
lat
long
1500000
1000000
500000
0
-500000
-1000000
1e+06
0e+00
-1e+06
-2e+06
-1e+06
0e+00
1e+06
2e+06
[0.242,0.373]
(0.373,0.401]
(0.401,0.426]
(0.426,0.451]
(0.451,0.479]
(0.479,0.508]
(0.508,0.545]
(0.545,0.587]
(0.587,0.778]

Notes: The figure plots average predicted income percentile of the son with a father in the 25th percentile across SEAs. US-born white males 30-36 and a father of 16-40 years older than the son, with valid variables (occupation, age, urban, race, literacy, county in SEAs) linked to fathers by census. Income percentiles constructed from occupations based on income data in 1940.

This page intentionally left blank

Bibliography

ABRAMITZKY, R., L. BOUSTAN, K. ERIKSSON, J. FEIGENBAUM, AND S. PÉREZ (2019): "Linking Individuals Across Historical Sources: a Fully Automated Approach," Working paper. 2

ABRAMITZKY, R., L. P. BOUSTAN, AND K. ERIKSSON (2012a): "Europe's Tired, Poor, Huddled Masses: Self-Selection and Economic Outcomes in the Age of Mass Migration," *American Economic Review*, 102, 1832–56. 2.1 3.2.1, 3.2.1.1

ABRAMITZKY, R., R. MILL, AND S. PÉREZ (2018): "Linking Individuals Across Historical Sources: a Fully Automated Approach," Working Paper 24324, National Bureau of Economic Research. 3.2.1, 3.2.1.2

ADAO, R., C. ARKOLAKIS, AND F. ESPOSITO (2019): "Spatial linkages, global shocks, and local labor markets: Theory and evidence," . 2.4.4

AHLFELDT, G. M., S. J. REDDING, D. M. STURM, AND N. WOLF (2012): "The Economics of Density: Evidence from the Berlin Wall," CEP Discussion Papers dp1154, Centre for Economic Performance, LSE. 1.1

AKCIGIT, U., J. GRIGSBY, AND T. NICHOLAS (2017): "Immigration and the Rise of American Ingenuity," Tech. rep., National Bureau of Economic Research. 2.1

ALLEN, T. AND C. ARKOLAKIS (2013): "Trade and the Topography of the Spatial Economy," NBER Working Papers 19181, National Bureau of Economic Research, Inc. 1.1

——— (2015): "A Model of Trade and Migration," Working paper. A.4

ALLEN, T., C. ARKOLAKIS, AND X. LI (2015): "Optimal City Structure," Working papers, Yale University. 1.1

ALLEN, T., C. D. C. DOBBIN, AND M. MORTEN (2018): "Border Walls," Working Paper 25267, National Bureau of Economic Research. A.4

ALONSO, W. (1964): *Location and land use. Toward a general theory of land rent.*, Cambridge, Mass.: Harvard Univ. Pr. 9

ARKOLAKIS, C., N. RAMONDO, A. RODRÍGUEZ-CLARE, AND S. YEAPLE (2018): "Innovation and production in the global economy," *American Economic Review*, 108, 2128–73. 2.4.2, 2.4.5

ATKESON, A. AND A. BURSTEIN (2010): "Innovation, Firm Dynamics, and International Trade," *Journal of Political Economy*, 118, 433–489. 2.4.1

BANERJEE, A., E. DUFLO, AND N. QIAN (2012): "On the Road: Access to Transportation Infrastructure and Economic Growth in China," NBER Working Papers 17897, National Bureau of Economic Research, Inc. 1.1

BAUM-SNOW, N. (2007): "Did Highways Cause Suburbanization?" Tech. Rep. 2. 1.1, 1.1

BECKER, G. AND N. TOMES (1986): "Human Capital and the Rise and Fall of Families," *Journal of Labor Economics*, 4, S1–39. 3.1

BOUSTAN, L. P. (2010): "Was Postwar Suburbanization "White Flight"? Evidence from the Black Migration," *The Quarterly Journal of Economics*, 125, 417–443. 1.1

CHETTY, R., N. HENDREN, P. KLINE, AND E. SAEZ (2014): "Where is the Land of Opportunity: The Geography of Intergenerational Mobility in the United States," *Quarterly Journal of Economics*, 129, 1553–1623, data available at&nbsp;http://www.equality-of-opportunity.org/index.php/data. 3.2.3, 3.4

DEKLE, R., J. EATON, AND S. KORTUM (2007): "Unbalanced trade," Tech. rep., National Bureau of Economic Research. 2.7

DONALDSON, D. (2010): "Railroads of the Raj: estimating the impact of transportation infrastructure," LSE Research Online Documents on Economics 38368, London School of Economics and Political Science, LSE Library. 1.1

DONALDSON, D. AND M. HORNBECK (2013): "Railroads and American Economic Growth: A &quot;Market Access&quot; Approach," NBER Working Papers 19213, National Bureau of Economic Research, Inc. 1.1

DURANTON, G., P. MORROW, AND M. TURNER (2013): "Roads and Trade: Evidence from the U.S," Working Papers tecipa-479, University of Toronto, Department of Economics. 1.1

ELLEN, I. G. AND K. O'REGAN (2011): "How low income neighborhoods change: Entry, exit, and enhancement," *Regional Science and Urban Economics*, 41, 89–97. 1.2.1.1

FABER, B. (2013): "Trade Integration, Market Size and Industrialization: Evidence from China's National Trunk Highway System," CEP Discussion Papers dp1244, Centre for Economic Performance, LSE. 1.1

Feigenbaum, J. J. (2015): "Intergenerational Mobility during the Great Depression," . 1.2.2, 3.1, 3.2.1, A.1.1

——— (2016): "Automated census record linking: A machine learning approach," . 6, 3.2.2

Ferrie, J. P. (1996): "A New Sample of Males Linked from the Public Use Microdata Sample of the 1850 U.S. Federal Census of Population to the 1860 U.S. Federal Census Manuscript Schedules," *Historical Methods: A Journal of Quantitative and Interdisciplinary History*, 29, 141–156. 3.2.1.1

Fujita, M. and H. Ogawa (1982): "Multiple equilibria and structural transition of non-monocentric urban configurations," Tech. Rep. 2. 1.1

Gilman, L. P. (1971): "The Development of a Neighborhood: Bedford, 1850-1880: A. Case Study," Master's thesis, columbia university. 13

Goeken, R., L. Huynh, T. Lynch, and R. Vick (2011): "New methods of census record linking," *Historical methods*, 44, 7–14. 1.2.2, 3.2.1, 3.2.2, A.1.1, A.1.1

Gordon, R. J. (2017): *The rise and fall of American growth: The US standard of living since the civil war*, Princeton University Press. 2.1

Heblich, S., S. Redding, and D. M. Sturm (2018): "The Making of the Modern Metropolis: Evidence from London," Working Paper 25047, National Bureau of Economic Research. 1.1

Jackson, K. T. (1985): *Crabgrass Frontier: The Suburbanization of the United States*, Oxford University Press, USA. 1.1, 1, 1.2.1.2, 1.4.2, 1.4.3, 13, 1.4.4.2, 1.5, 3

Kelly, B., D. Papanikolaou, A. Seru, and M. Taddy (2018): "Measuring technological innovation over the long run," Tech. rep., National Bureau of Economic Research. 2.3.1.4

Kortum, S. (1997): "Research, Patenting, and Technological Change," *Econometrica*, 65, 1389–1419. 2.1

Long, J. and J. Ferrie (2013): "Intergenerational Occupational Mobility in Great Britain and the United States since 1850," *American Economic Review*, 103, 1109–37. 3.2.1.1

Lucas, R. E. and E. Rossi-Hansberg (2002): "On the Internal Structure of Cities," Tech. Rep. 4. 1.1

Lucas Jr, R. E. and B. Moll (2014): "Knowledge growth and the allocation of time," *Journal of Political Economy*, 122, 1–51. 2.1

Mazumder, B. (2005): "Fortunate Sons: New Estimates of Intergenerational Mobility in the United States Using Social Security Earnings Data," *The Review of Economics and Statistics*, 87, 235–255. 3.1

Michaels, G. (2008): "The Effect of Trade on the Demand for Skill: Evidence from the Interstate Highway System," *The Review of Economics and Statistics*, 90, 683–701. 1.1

Mills, E. S. (1967): "An Aggregative Model of Resource Allocation in a Metropolitan Area," *The American Economic Review*, 57, 197–210. 9

Monte, F., S. J. Redding, and E. Rossi-Hansberg (2015): "Commuting, Migration and Local Employment Elasticities," CEP Discussion Papers dp1385, Centre for Economic Performance, LSE. 1.1

Muth, R. F. (1964): *Cities and Housing: The Spatial Pattern of Urban Residential Land Use.*, The University of Chicago Press. 9

Nix, E. and N. Qian (Working Paper): "The Fluidity of Race: "Passing" in the United States, 1880-1940," . 3.2.1

Nunn, N., N. Qian, and S. Sequeira (2017): "Migrants and the Making of America: The Short and Long Run Effects of Immigration During the Age of Mass Migration," . 2.1

Olivetti, C. and M. D. Paserman (2015): "In the name of the son (and the daughter): Intergenerational mobility in the united states, 1850-1940," *American Economic Review*, 105, 2695–2724. 3.1, 3.2.3, 3.2.3

Perla, J. and C. Tonetti (2014): "Equilibrium imitation and growth," *Journal of Political Economy*, 122, 52–76. 2.1

Preston, S. H. and M. R. Haines (1991): *Fatal Years: Child Mortality in Late Nineteenth-Century America*, Princeton University Press. 3.2.3

Roca, J. D. L. and D. Puga (2017): "Learning by working in big cities," *The Review of Economic Studies*, 84, 106–142. 3

Romer, P. M. (1990): "Endogenous Technological Change," *The Journal of Political Economy*, 98, S71–S102. 2.1

Ruggles, S., S. Flood, R. Goeken, J. Grover, E. Meyer, J. Pacas, and M. Sobek (2019): "IPUMS USA: Version 9.0 [dataset]," . 1.2.1.1

Shertzer, A., R. P. Walsh, and J. R. Logan (2016): "Segregation and Neighborhood Change in Northern Cities: New Historical GIS Data from 1900 to 1930," Tech. rep. 1.2.3

Solon, G. (1992): "Intergenerational Income Mobility in the United States," *The American Economic Review*, 82, 393–408. 3.1

Tsivanidis, N. (2018): "The Aggregate And Distributional Effects Of Urban Transit Infrastructure: Evidence From Bogotá's TransMilenio," . 1.1

Appendix A

Appendix for Chapter 1

Appendix to Chapter 1

In this section, I describe the record linking procedure and relevant details. In constructing a panel of individuals, I use "Machine Learning," where the machine can learn the pattern of "true" and "false" matches and self-link individuals after learning the patterns of true and false matches from training datasets. This method is implemented to link individuals across census years while maximizing the match rate and representativeness of linked datasets. I link complete-count US Federal Decennial Demographic Census records from 1850 to 1940 with newly transcribed socioeconomic variables such as occupation and industry.

A.1 Methodology

A.1.1 Machine Learning Approach of Record Matching

The "machine learning" approach for record linking borrows insights from computer science and statistics and I implement this method of classification and text comparison to link individual records. The rationale behind my choice of machine learning is to learn from big data. In essence, record linking without unique identifier is to predict whether certain linked records are "true" links of the same individual or not, based on a set of features such as first name and last name, age, and place of birth. Similar efforts have been pioneered by Goeken et al. (2011) that create the IPUMS linked samples. Feigenbaum (2015) links individual records of the 1915 Iowa State Census to their adult-selves in the 1940 US Federal Demographic Census records. Relative to the mentioned work, my record linking is far more extensive in the scope of matching as this involves complete-count US Federal Decennial Demographic Census records of all years from 1850 to 1940. I teach a machine to learn to predict based on a set of features. I create a training dataset in which contain both "true" and "false" matches and their characteristics (e.g some observations with "true" as an outcome would have same/very similar characteristics in terms of age, first and last name, parents' and his/her birthplaces whereas observations with "false" as an outcome would have quite different characteristics in terms of the above mentioned characteristics). In this case, the outcome is whether the matched records are "true" or "false" match, given the observed characteristics. By taking this training data, I build a prediction model, or learner, which will enable us to predict the outcome for new, unseen objects. A well-designed learner armed with a solid training dataset should accurately predict outcomes for new unseen objects.

I implement a supervised learning problem in the sense that the presence of outcome variable ("true" or "false" links) guides the learning process—in other words, the end-goal is to use the inputs to predict the output values. To summarize this process, I extract subsets of possible matches for each record and create training data in order to tune a matching algorithm so that the matching algorithm matches individual records by minimizing both false positives and false negatives while reflecting inherent noises in historical records. I have explored various models for model selection. By comparing and analyzing matched records that I match through various methods, I choose the random forest classification as it is *more conservative* in matching records—the number of matched records is lower than that of Support Vector Machine (hereafter, SVM)— and the number of unique matches are significantly higher than the standard SVM model. Although the choice of random forest classification may result in lower number match rate due to its conservative nature, I integrated household-level information in linking individual records to mitigate the concerns of low match rate.

A filtering process called "pruning" for non-unique matches

Although I largely follow the standard machine-learning record linking methodology suggested by Goeken et al. (2011), I have extended the techniques of Goeken et al. (2011) by inventing a two-step machine learning matching methodology. Especially, I make use of the parents and/or spouse information such as birthplaces and names to choose the "true" match among other candidate matches. This additional step of extracting household-level information and its use in selecting "true" matches among multiple candidates (instead of dropping non-unique matches, which have been the "standard" practices in the existing record matching literature) is novel. This procedure can not only save a number of matches that otherwise had to be dropped but also correct for the selection bias (people with common characteristics such as common first and last names may be systematically under-represented in linked datasets).

A.1.2 Record Linking in Practice: Innovations

The core of census matching is a classification problem. Given any pair of records from different census years, finding a true match is to find the mapping that classifies the pair as matched or unmatched based on the set of pre-determined features, including names, gender, age, race and birthplace. However, since this set of features is far from unique, there are cases where one individual has several candidate matches (e.g. there are many "John Smith" with same age).

Most record linking approaches throw away non-unique matches. One of the contributions of my record linking approach is the use of household-level information to turn the non-unique types of matches (second to fourth type) as unique matches. Specifically, I use father and mother's information such as their racial background, birthplaces, birth year) and use the same information for spouses of individuals. This not only increases the match rates but also alleviates the concern of systematic selection bias (e.g. people with common given given and last names may be systematically under-represented in the linked data).

A.1.2.1 Female Record Linking

Historically, as women typically change their last names after marriage (and in the absence of time-invariant individual-level unique identifiers such as social security number in historical records), female record linking has been challenging. To my knowledge, this is one of the first endeavors of linking historical female records. I assume that women's last names are likely to change if their marital status changes from single in the earlier period to married in the later period, and I do not consider record linking for such case. On the other hand, I assume one's last name is likely to remain the same if the marital status is either married to the same partner in both years, or married in the early period and then widowed or divorced in the later period; or remained single in both periods.

A.2 The Transportation Revolution

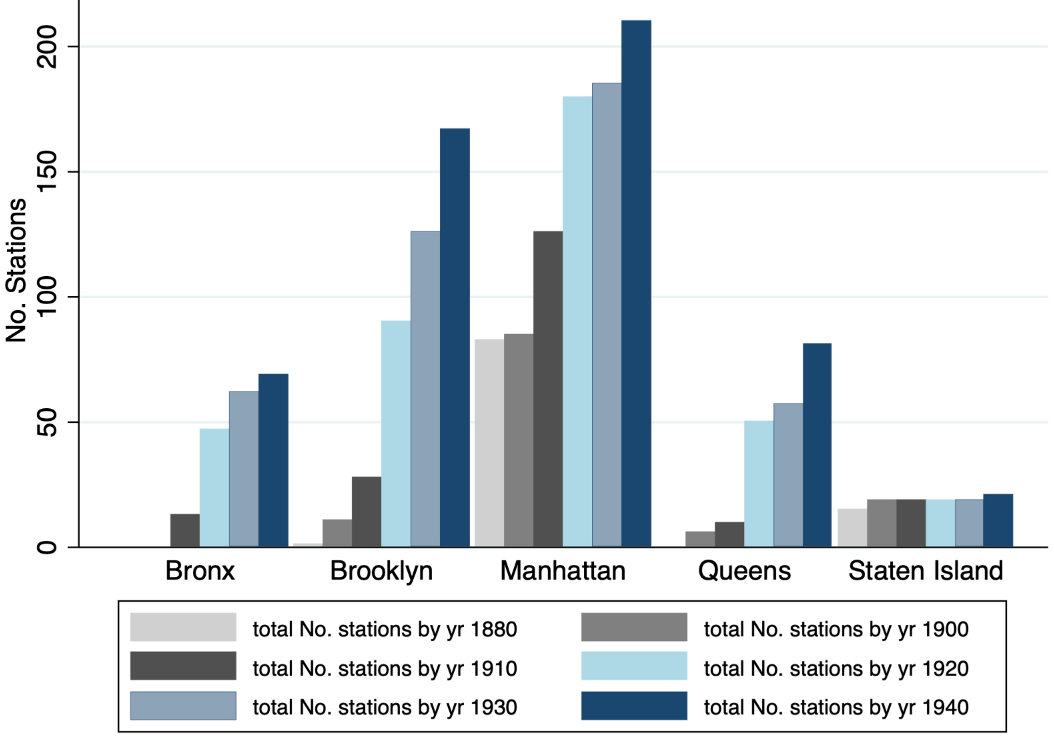
Transit infrastructure improved dramatically at both intra- and inter-city level during the study period. Figure A1 shows the total number of intra-city railways and subway stations by borough by the end of each decade during the study period. Especially, during the subway construction period between 1904 and 1920, the total number of stations grew by 200% and 113% in the Bronx, 87% and 105% in Brooklyn, 50% and 133% in Queens.

Inter-city transit infrastructure improvements at an unprecedented scale during the study period as well: electrification of railroads that served central Westchester county, Connecticut in 1907 and 1914 improved the efficiency and speed of railways greatly, the Hudson Tubes that connected New Jersey was built in 1908, and inter-city railway that connected NYC to the rest of the country with the opening of Penn Station in 1910.

A.2.1 Intra-city transit infrastructure changes

- Subways and Elevated Railways Construction and Network Change over Time

Figure A1: Total number of intra-city railway, subway stations



Source: Author's creation using New York Transit Museum Archive.

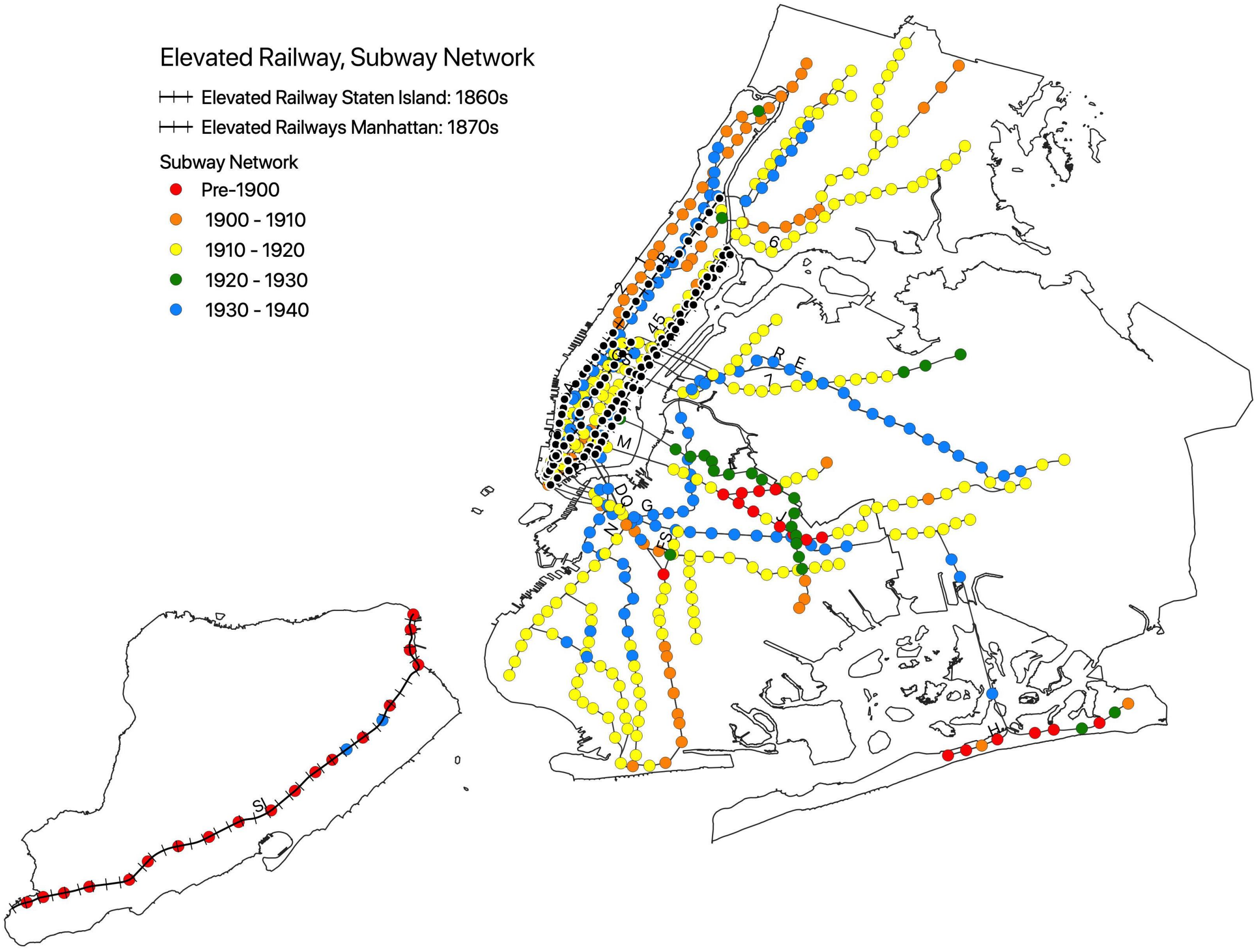
Figure A2 captures the evolution of spatial links by intra-city commuting transit infrastructure which are the elevated and subterranean railways. Before the introduction of the subway in 1904, New York City had a large central business district in lower Manhattan and a smaller business district in downtown Brooklyn. These districts were served by elevated railways and ferries and most of the services were operating in Manhattan. As Figure A2 shows, elevated lines ran north from the southern tip of Manhattan to the Bronx. There were very few east-west connections in Manhattan and this pattern persisted for the subway network in the twentieth century as well. Before the introduction of the subway in 1904, Manhattan was the only borough with rapid mass transit commuting infrastructure. Most outer boroughs (i.e. Queens, Staten Island, and the Bronx) did not have transit network into the 1910s and were semi-rural and underdeveloped. Figure A1 shows the total number of stations by the borough over time. The first decade of subway construction mostly served Manhattan and Brooklyn, whereas parts of Bronx, Queens and South Brooklyn received more subway constructions in the 1910s under the Dual Contracts. However, the rapid growth of the system largely was over by 1940.[1]

- Intra-city transit access measures by the elevated railways and subways

[1] The first underground line of the subway opened in 1904, almost 40 years after the opening of the first elevated railway in Manhattan. New York City subway was built by two private companies (the Brooklyn Rapid Transit Company (BRT, later Brooklyn–Manhattan Transit Corporation, BMT) and the Interborough Rapid Transit Company (IRT)) and one city-owned company (Independent Subway System (IND)). In 1940, the city bought the two private systems and consolidated the transit network.

I define Transit Access (hereafter, TA) as the number of stations in each neighborhood.[2] The number of total stations as a proxy for transit access is convenient in understanding a form of hub-spoke distribution paradigm where a series of "spokes" that connect outlying points to a central "hub." Before the introduction of subways, lower Manhattan ("Downtown Manhattan") was the area where the transit network is extremely well connected ("transit hubs"). However, as subway expanded and inter-city transit infrastructure was largely built in Midtown Manhattan, "transit hubs" expanded from Downtown Manhattan to Midtown Manhattan. I describe the spatial links by inter-city transit infrastructure in the following Subsubsection A.2.2.

Figure A2: Evolution of Spatial Links by the Elevated Railway, Subways

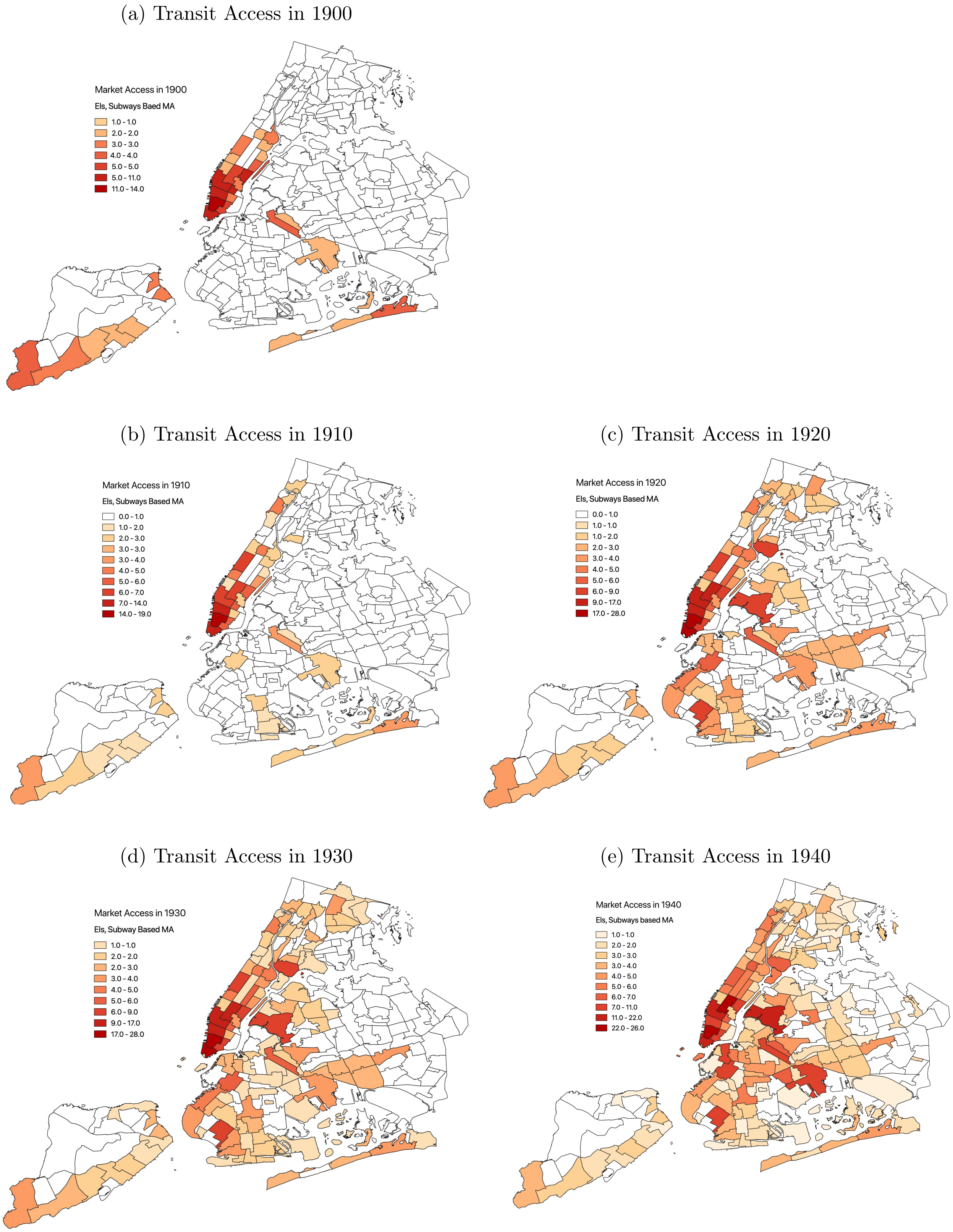

Elevated Railway, Subway Network
Elevated Railway Staten Island: 1860s
Elevated Railways Manhattan: 1870s
Subway Network
Pre-1900
1900 - 1910
1910 - 1920
1920 - 1930
1930 - 1940
6
R
E
7
M
D
Q
G
N
F
S
SI
45
2
B
A
H

Note: The above figures show the evolution of intra-city spatial links in terms of the elevated railway, subways over study period. Different colors denote the opening years of transit links. Source: Author's Creation using New York City Department of City Planning's data called "LION" GIS data which is a base map representing the city's geographic features.

- Transit access changes based on the elevated railway and subway

[2] In the Appendix, I map transit access change over the study period based on intra-city mass transit infrastructure (i.e. the elevated railways and subways)

Figure A3: The Els, Subways-Based Transit Access Measures by Decade



(a) Transit Access in 1900

(b) Transit Access in 1910

(c) Transit Access in 1920

(d) Transit Access in 1930

(e) Transit Access in 1940

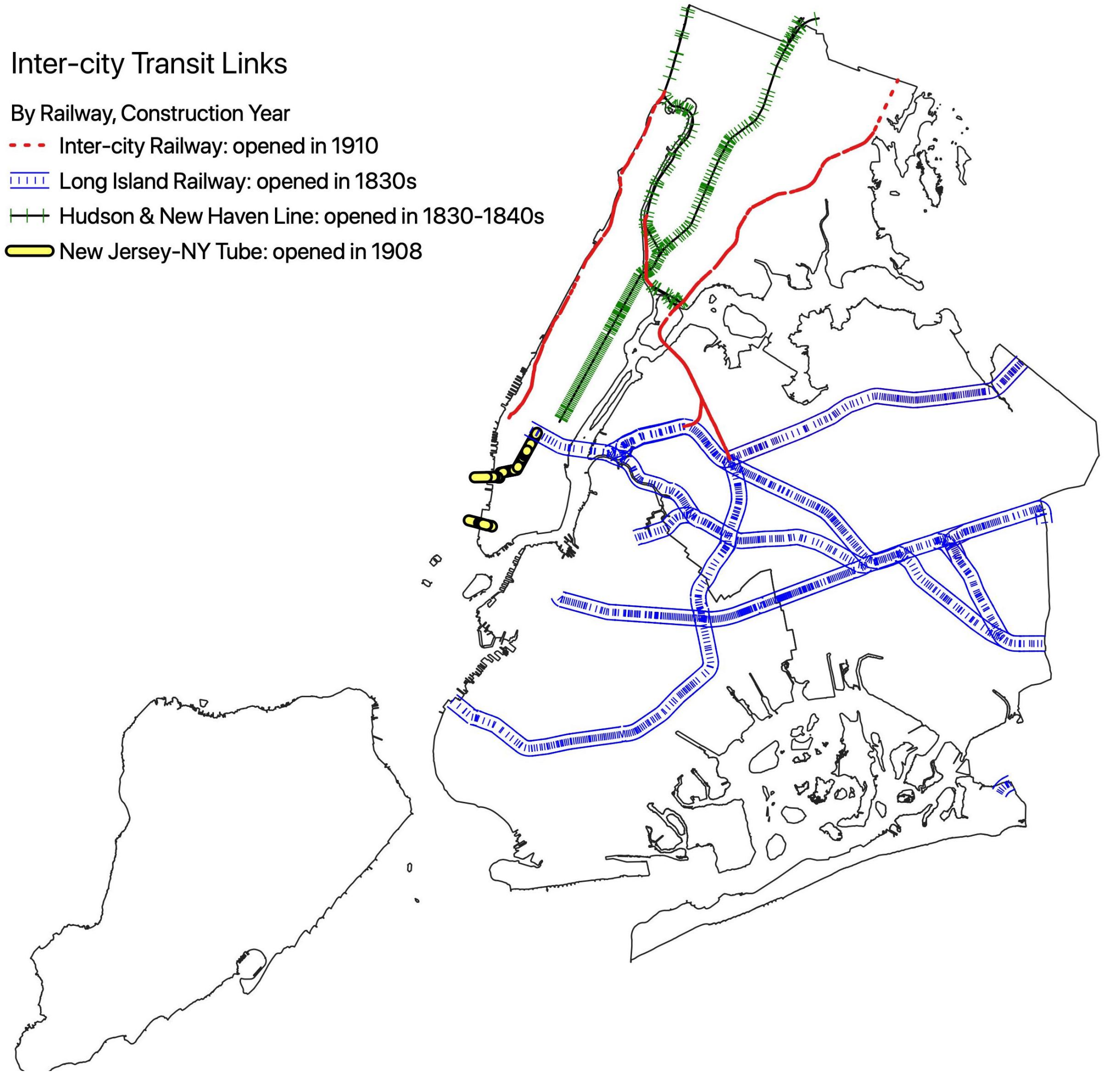
Note: The above figures show transit access by decade based on intra-city mass transit infrastructure (i.e. the elevated railways and subways).

A.2.2 Inter-city transit infrastructure changes

Inter-city transit infrastructure was largely concentrated in Midtown Manhattan, and the combination of both inter- and intra-city transit infrastructure improvements grew faster in Midtown than in Lower Manhattan. By the Year 1910, the inter-city transit infrastructure-based transit access in NYC experienced an unprecedented, spectacular growth—steam railroad began in the 1830s by New York and Harlem Railroad (Green line); by the 1840s, the same line served central Westchester county; Long Island Railroad (LIRR)-based commuter service was established largely by the 1860s (Blue line); the Hudson Tubes, which became Port Authority Trans-Hudson (PATH) opened in 1908 (Yellow line), and inter-city railway connected NYC to the rest of the country with the opening of Penn Station in 1910 (Red line). Figure A.2.2 shows the spatial pattern of inter-city transit infrastructure improvements over the study period. NYC's transit "hubs" expanded from Downtown Manhattan to Midtown Manhattan and this change and the extreme growth of Midtown Manhattan was partly due to inter-city railway infrastructure that connects the NYC's surrounding regions.[3]

[3] Jackson (1985) argues the first railroads were designed for long-distance rather than local travel. However, as railroad companies sought revenues, they built stations whenever their lines passed through rural villages on the outskirts of larger cities. Jackson (1985) argues that as inter-city railway fares were considered too high for most wage earners, such suburbanization was only for the "well-to-do."

Figure A4: Evolution of Inter-City Transit Infrastructure by Construction Year

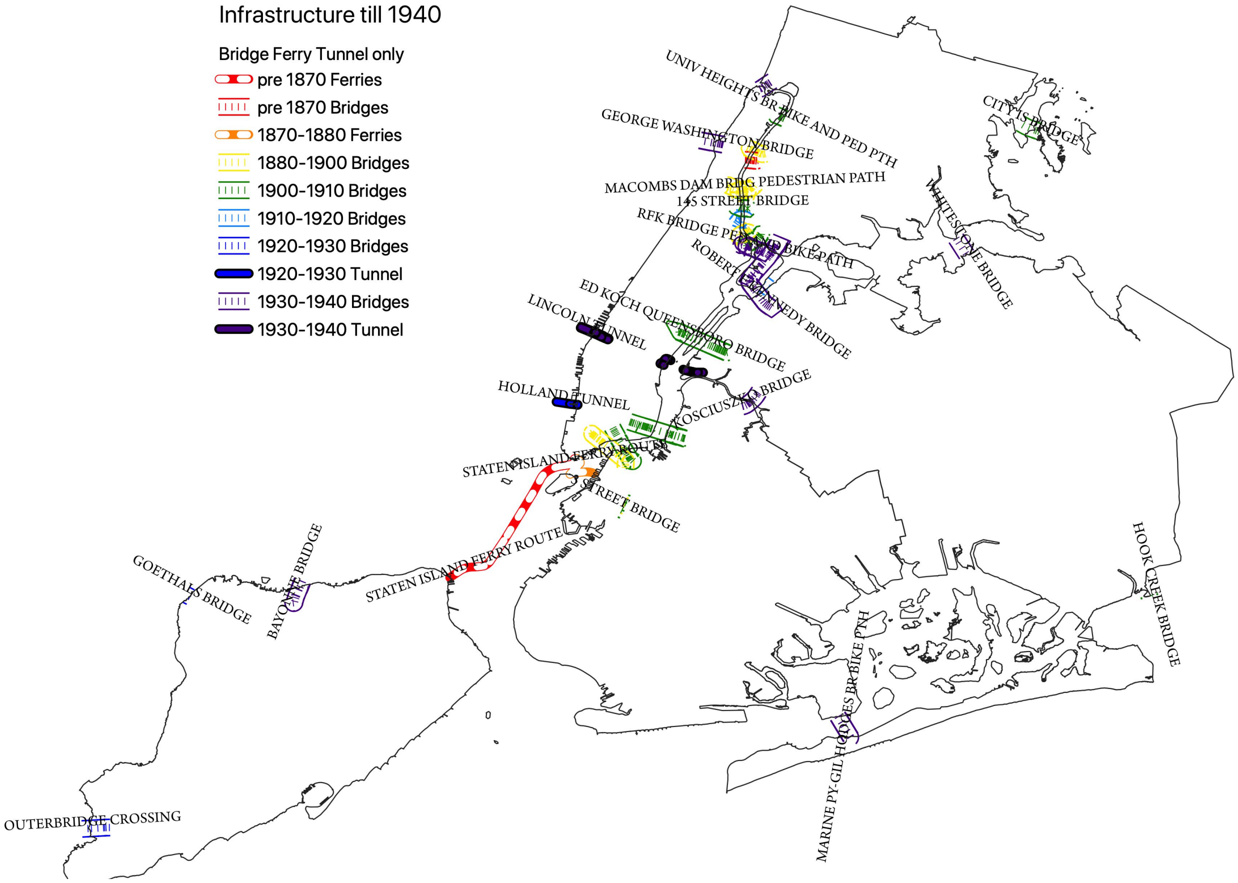

Inter-city Transit Links
By Railway, Construction Year
Inter-city Railway: opened in 1910
Long Island Railway: opened in 1830s
Hudson & New Haven Line: opened in 1830-1840s
New Jersey-NY Tube: opened in 1908

Note: The above figure shows the evolution of inter-city transit networks by railroad network and construction year. I construct the following information based on information provided by the New York City Transit Authority and related books (http://www.mta.info/).

- Bridges, Ferries, and Tunnels

Although the railway is my primary focus of the study, water-borne transportation played an important role in forming connections between the core and connecting regions such as Brooklyn, Staten Island, and parts of New Jersey. As the city economy depended on water-borne transport, extensive bridge-building followed: the Brooklyn Bridge (1883), Williamsburg Bridge (1903), Manhattan Bridge (1909), and Queensboro Bridge (1909) were constructed over the East River. The Hell Gate Bridge (1917) carried trains of the Pennsylvania Railroad and finally, George Washington Bridge (1931) connected New Jersey and New York City. In the Appendix, I map bridges, ferries, and tunnels that were constructed during the study period to connect boroughs in the City.

Figure A5: Evolution of Spatial Links by Bridges, Ferry, Tunnel



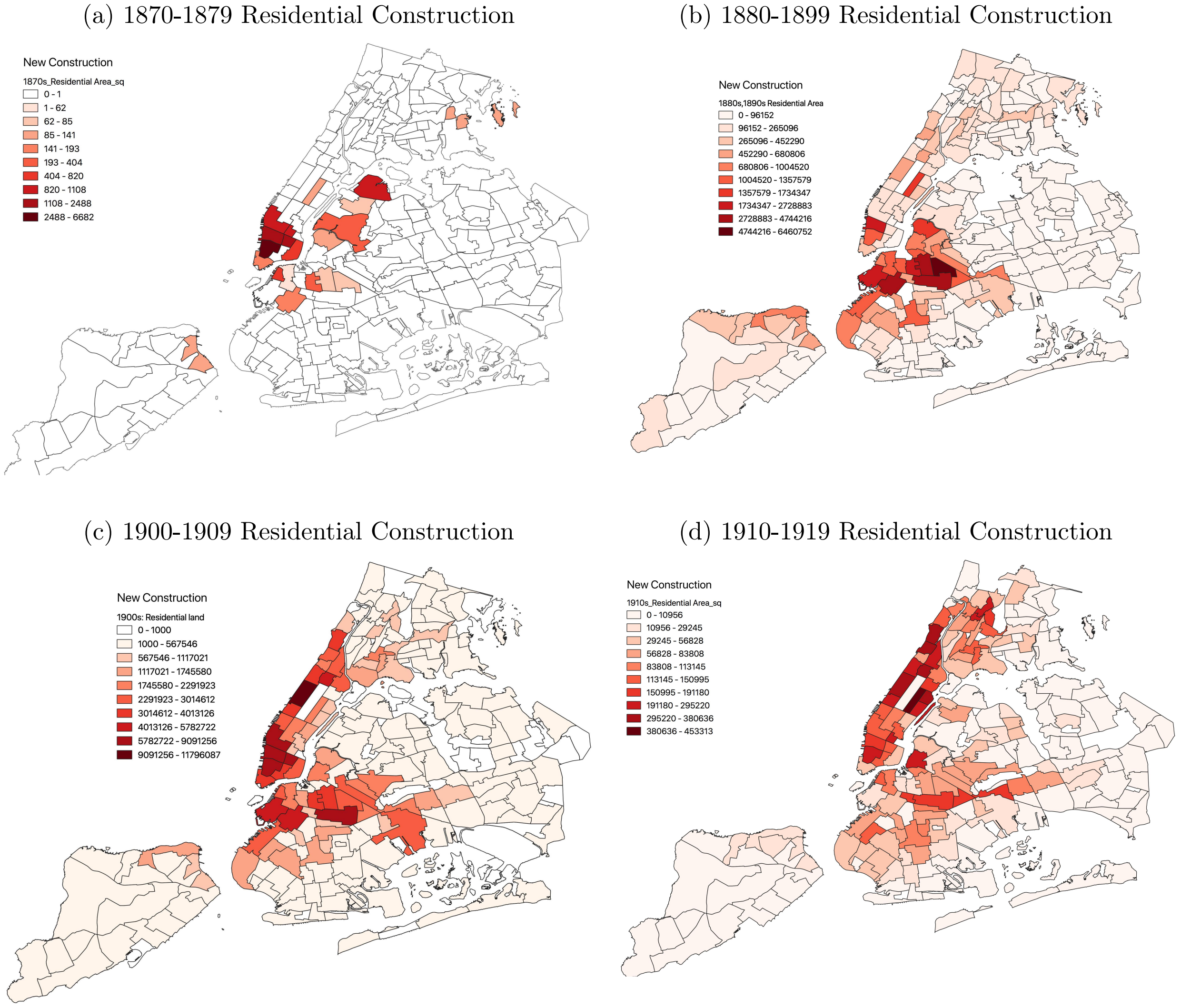
Note: The above figures show the evolution of intra-city spatial links in terms of bridges, tunnels, and ferries between census periods. Different colors denote the opening years of transit links. Source: Author's Creation using New York City Department of City Planning's data called "LION" GIS data which is a base map representing the city's geographic features.

A.3 Supplementary Figures

A.3.1 Different Land Use Creation

A.3.1.1 Residential Land Use Construction

Figure A6: New Construction of Residential-Use Land



(a) 1870-1879 Residential Construction

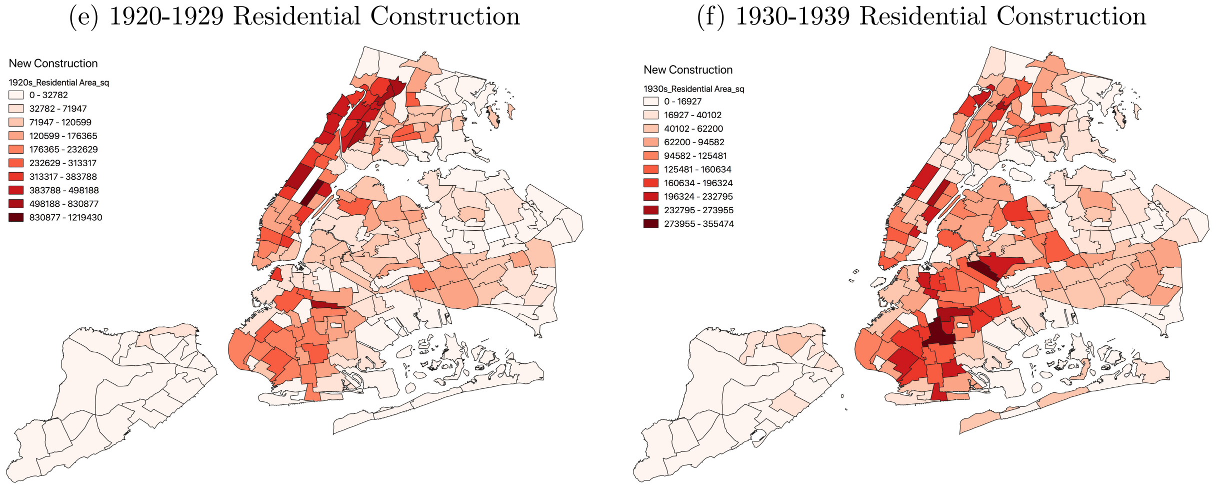
(b) 1880-1899 Residential Construction

(c) 1900-1909 Residential Construction

(d) 1910-1919 Residential Construction

Note: The above figures show percent change of population density between two adjacent census periods. Source: Author's Creation using the complete-count US Federal Demographic Census from 1870 to 1940.

Figure A6: New Construction of Residential-Use Land



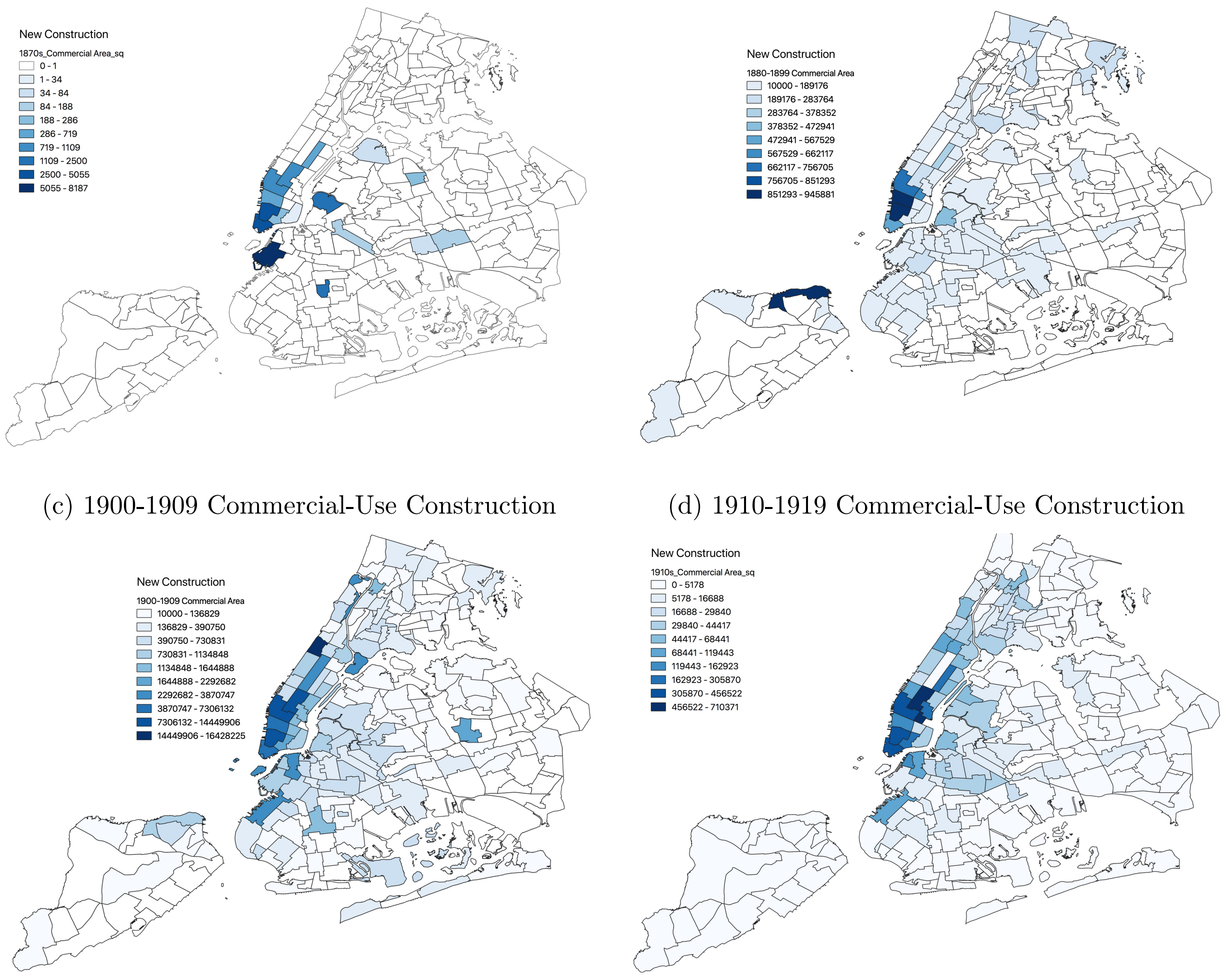
(e) 1920-1929 Residential Construction

(f) 1930-1939 Residential Construction

Note: The above figures show percent change of population density between two adjacent census periods. Source: Author's Creation using the complete-count US Federal Demographic Census from 1870 to 1940.

A.3.1.2 Commercial Land Use Construction

Figure A7: New Construction of Commercial-Use Land



(a) 1870-1879 Commercial-Use Construction

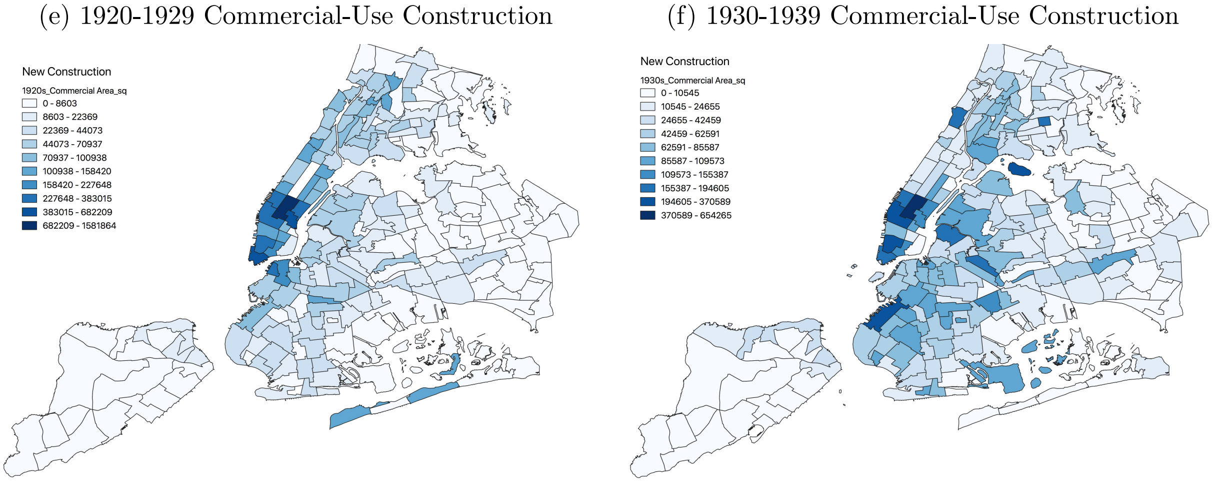
(b) 1880-1899 Commercial-Use Construction

(c) 1900-1909 Commercial-Use Construction

(d) 1910-1919 Commercial-Use Construction

Note: The above figures show percent change of population density between two adjacent census periods. Source: Author's Creation using the complete-count US Federal Demographic Census from 1870 to 1940.

Figure A7: New Construction of Commercial-Use Land


(e) 1920-1929 Commercial-Use Construction

(f) 1930-1939 Commercial-Use Construction

Note: The above figures show percent change of population density between two adjacent census periods. Source: Author's Creation using the complete-count US Federal Demographic Census from 1870 to 1940.

A.4 Theoretical framework

In this section, I present the theoretical framework based on Allen and Arkolakis (2015), and Allen et al. (2018). This general equilibrium spatial framework features a dynamic setting where workers differing in skill and nativity choose where and how to migrate between different locations. This framework allows me to assess the welfare effects of transit-infrastructure driven market access improvement on workers with different nativity and skill in different locations.

A.4.1 Setup

Geography

There is a world comprised of a compact set $i \in \{1, ..., N\} \equiv \mathcal{N}$ of locations and inhabited by workers with different nativity n (foreign-born F, native-born U) and skills s(high skill h and low-skill l), each endowed with a unit of labor which they supply inelastically. Let $L_{it}^{n,s}$ denote the number of workers in location i of nativity n and skill s. In each location $i \in N$, the four type of workers combine their labor to produce a differentiated variety of good using a nested constant elasticity of substitution (CES) production function:

$$Q_i = (\sum_{s\in\{h,l\}} ((\sum_{n\in\{F,U\}} A_i^{n,s}(L_i^{n,s})^{\frac{\rho_s-1)}{\rho_s}})^{\frac{\rho_s}{\rho_s-1}}))^{\frac{\rho-1}{\rho}})^{\frac{\rho}{\rho-1}} \tag{A.1}$$

where $A_i^{n,s} > 0$ is the productivity of a worker of nativity n and skill s in location i, $\rho_s \geq 1$ is the elasticity of substitution across the nativity of workers of a skill s, and $\rho \geq 1$ is the elasticity of substitution across high-skill and low-skill workers.

Production

Workers in location i with (composite) productivity $A_i^{n,s} > 0$ earn an (endogenous) wage $w_i^{n,s}$. Product markets are perfectly competitive and a worker in location i of nativity n and skill $s$is paid a wage $w_i^{n,s}$ equal to her marginal product:

$$w_i^{n,s} = p_i \times (Q_i)^{\frac{1}{\rho}} \times ((\sum_{n \in \{F,U\}} A_i^{n,s} (L_i^{n,s})^{\frac{\rho_s - 1}{\rho_s}})^{\frac{\rho_s}{\rho_s - 1}})^{(\frac{1}{\rho_s} - \frac{1}{\rho})} \times A_i^{n,s} \times (L_i^{n,s})^{-\frac{1}{\rho_s}} \tag{A.2}$$

, where p_i is the equilibrium price of the differentiated variety produced in location i. Under perfect competition and production function above, $p_i$takes the following form:

$$p_i = (\sum_{s \in \{h,l\}} ((\sum_{n \in \{F,U\}} (A_i^{n,s})^{\rho_s} (w_i^{n,s})^{1-\rho_s})^{\frac{1}{1-\rho_s}})^{1-\rho})^{\frac{1}{1-\rho}} \tag{A.3}$$

where $P_i \equiv (\sum_{j \in N} (\tau_{ij} p_j)^{1-\sigma})^{\frac{1}{1-\sigma}}$ is the Dixit-Stiglitz price index, and $u_i^{n,s}$ is a type-specific amenity for each location.

Trade

As workers have CES preferences over varieties and each location produces a differentiated variety, workers will consume varieties produced in other locations. We assume that trade between locations is subject to "iceberg" trade costs such that $\tau_{ij} \geq 1$ units of a good produced in location $i \in S$ must be shipped in order for one unit to arrive in location $j \in \mathcal{N}$; As a result, the price of a differentiated variety from a location $i \in \mathcal{N}$ and in location $j \in \mathcal{N}$ is $p_{ij} = \tau_{ij} p_i$. Workers have CES preferences over varieties produced in all locations with elasticity of substitution $\sigma \geq 1$ and their indirect utility can be written as

$$W_i^{n,s} = \frac{w_i^{n,s}}{P_i} u_i^{n,s} \tag{A.4}$$

Given the setup of iceberg trade costs and perfect competition in the production market, gravity trade equation of the value of trade from location $i \in \mathcal{N}$ to location $j \in \mathcal{N}$, X_{ij}, can be written as:

$$X_{ij} = \tau_{ij}^{1-\sigma} (p_i)^{1-\sigma} P_j^{\sigma-1} E_j, \tag{A.5}$$

where E_j is the total expenditure in location j.

A.4.2 Migration

A.4.2.1 Migration decision on which labor market to face

The movement of people across locations are also subject to "iceberg" frictions. For simplicity, we take the initial distribution of heterogenous workers with different nativity and skill $\{L_{i,0}^{n,s}\}$

as exogenous and treat the migration decision as static. Then, a continuum of heterogenous workers $\nu \in [0, L_{i,0}^{n,s}]$ choose where to live in order to maximizer her welfare:

$$U_i^{n,s}(\nu) = \max_{j \in \mathcal{N}} \times \frac{W_j^{n,s}}{\mu_{ij}^{n,s}} \times \varepsilon_{ij}^{n,s}(\nu), \tag{A.6}$$

where $\mu_{ij}^{n,s} \geq 1$ is a migration friction common to all workers moving from location $i \in \mathcal{N}$ to location $j \in \mathcal{N}$ of type $\{n,s\}$, and $\varepsilon_{ij}^{n,s}(\nu)$ is a migration friction idiosyncratic to workers νdrawn from an extreme value (Fréchet) distribution with shape parameter $\theta^{n,s} \geq 0$. We assume that amenity of a particular place depends on an exogenous term and the local population:

$$L_{ij}^{n,s} = \left(\mu_{ij}^{n,s}\right)^{-\theta^{n,s}} \left(\frac{w_j^{n,s}}{P_j} u_j^{n,s}\right)^{\theta^{n,s}} \left(\Pi_i^{n,s}\right)^{-\theta^{n,s}} \left(L_{i,0}^{n,s}\right), \tag{A.7}$$

where $\left(\Pi_i^{n,s}\right) = \left(\sum_{j \in \mathcal{N}} \left(\mu_{ij}^{n,s}\right)^{-\theta^{n,s}} \left(W_j^{n,s}\right)^{\theta^{n,s}}\right)^{\frac{1}{\theta^{n,s}}}$.

Equation A.7 is a gravity equation for migration: all else equal, there will be greater flows from location $i \in \mathcal{N}$ to location $j \in \mathcal{N}$ of type $\{n,s\}$ the lower bilateral migration costs of workers with nativity n and skills s, $\mu_{ij}^{n,s}$, the higher type-specific amenity in location $j \in \mathcal{N}$ for workers with nativity and skill pair $\{n,s\}$, $u_j^{n,s}$, the higher real wages in location $j \in \mathcal{N}$ for workers with nativity and skill pair $\{n,s\}$, $\frac{w_j^{n,s}}{P_j}$.

A.4.2.2 Neighborhood decision and commuting costs

Suppose now that the heterogenous workers with different nativity and skill choose which neighborhood to live ($k \in \mathcal{K}$), conditional on working in region j. All neighborhoods $k \in K$ are regions such that commuting to location j is feasible, with commuting cost κ_{jk}. Under the Cobb-Douglas preferences, worker preferences are defined over consumption goods and residential floor space, with the indirect utility for a worker $\nu \in [0, L_{i,0}^{n,s}]$ residing in ($k \in \mathcal{K}$), working in j is:

$$U_{jk}^{n,s}(\nu) = \max_{k \in \mathcal{K}} \times \frac{w_j^{n,s} u_k^{n,s}}{\kappa_{jk} P_k Q_k} \times \varepsilon_{jk}^{n,s}(\nu), \tag{A.8}$$

$$W_j^{n,s} = E_K U_{jk}^{n,s}(\nu) = \left[\sum_{k \in \mathcal{K}} \left(\frac{w_j^{n,s} u_k^{n,s}}{\kappa_{jk} P_k Q_k}\right)^{\Theta}\right]^{\frac{1}{\Theta}}$$

P_k is the price index for consumption goods; Q_k is the price of floor space, $w_j^{n,s}$is the wage of workers with nativity and skill $\{n,s\}$,κ_{jk} is an iceberg commuting cost between region j and neighborhood $k \in K$, and commuting costs are same across workers with different nativity and skill, and $\varepsilon_{jk}^{n,s}(\nu)$ is an idiosyncratic amenity draw that captures all the idiosyncratic factors that cause an individual to live and work in particular locations within the city, and $\varepsilon_{jk}^{n,s}(\nu)$ is a commuting friction idiosyncratic to workers νdrawn from an extreme value (Fréchet) distribution with shape parameter $\Theta \geq 0$.

where $(\Pi_i^{n,s}) = \left(\sum_{j\in\mathcal{N}} \left(\mu_{ij}^{n,s}\right)^{-\theta^{n,s}} \left(W_j^{n,s}\right)^{\theta^{n,s}}\right)^{\frac{1}{\theta^{n,s}}}$.

A.4.3 Equilibrium

Given a geography of the world, the model elasticities, and the initial distribution of population $\{L_{i,0}^{n,s}\}$, the equilibrium of the model is defined by a set of location observables such that:

1. (Law of Motion of Migration) Given wages and the price index, the number of workers of different nativity n (foreign-born F, native-born U) and skills s in each location is equal to the total flows of workers to that location:

$$L_{ij}^{n,s} = \left(\mu_{ij}^{n,s}\right)^{-\theta^{n,s}} \left(\frac{w_j^{n,s}}{P_j} u_j^{n,s}\right)^{\theta^{n,s}} \left(\Pi_i^{n,s}\right)^{-\theta^{n,s}} \left(L_{i,0}^{n,s}\right),$$

2. Given the number of workers in each location, the quantity of produced of the differentiated variety in each location takes the production function from Equation A.1

3. (labor market clearing) Given the number of workers in each location, the equilibrium price and quantity produced of the differentiated variety, the equilibrium wage of each type worker with nativity and skill pair $\{n, s\}$ in each location is equal to its marginal product, as in Equation A.2

4. (balanced trade) Given the quantity produced of the differentiated variety in each location, equilibrium prices are determined by the income and expenditure of a location being equal to its total sales:

$$p_i Q_i = \sum_{j\in\mathcal{N}} \tau_{ij}^{1-\theta} p_i^{1-\theta} P_j^{\sigma-1} p_j Q_j$$

This page intentionally left blank

Appendix B

Appendix for Chapter 2

Appendix to Chapter 2

B.1 Proofs

B.1.1 Proof of Proposition 2.4.1

Consider the value function in (2.9), given by

$$V_{rt}\left(q\right)=\pi_{rt}\left(q\right)+\max_{i_t}\left[\frac{1}{1+r_t}V_{rt+1}\left(qi_t\right)-w_{rt}^{R}\frac{1}{\zeta_r^I Z_t^I}\frac{q}{Q_{rt}^{\lambda}}\frac{i_t^{\iota}}{\iota}\right].$$

We conjecture that $V_{rt}\left(q\right)$ is linear in q, i.e. takes the form

$$V_{rt}\left(q\right)=\upsilon_{rt}q, \tag{B.1}$$

and we will determine υ_{rt}. Using, (B.1) and (2.8) we get that

$$\upsilon_{rt}q=\frac{1}{\sigma}\frac{q}{Q_{rt}}\frac{E_{rt}}{N_{rt}}+\max_{i_t}\left[\frac{1}{1+r_t}\upsilon_{rt+1}iq-w_{rt}^{R}\frac{1}{\zeta_r^I Z_t^I}\frac{q}{Q_{rt}^{\lambda}}\frac{i_t^{\iota}}{\iota}\right],$$

so that indeed

$$\upsilon_{rt}=\frac{1}{\sigma}\frac{1}{Q_{rt}}\frac{E_{rt}}{N_{rt}}+\max_{i_t}\left[\frac{1}{1+r_t}\upsilon_{rt+1}i-w_{rt}^{R}\frac{1}{\zeta_r^I Z_t^I}\frac{1}{Q_{rt}^{\lambda}}\frac{i_t^{\zeta}}{\zeta}\right].$$

The optimality condition for i_t reads

$$\frac{1}{1+r_t}\upsilon_{rt+1}=w_{rt}^{R}\frac{1}{\zeta_r^I Z_t^I}\frac{1}{Q_{rt}^{\lambda}}i_t^{\iota-1}.$$

This implies that the optimal innovation rate is given by

$$i_t=\left(\frac{1}{1+r_t}\frac{\upsilon_{rt+1}}{w_{rt}^{R}}Q_{rt}^{\lambda}\zeta_r^I Z_t^I\right)^{\frac{1}{\iota-1}}, \tag{B.2}$$

and that the value function is given by

$$\upsilon_{rt}=\frac{1}{\sigma}\frac{1}{Q_{rt}}\frac{E_{rt}}{N_{rt}}+\left(\frac{\iota-1}{\iota}\right)w_{rt}^{R}\frac{1}{\zeta_r^I Z_t^I}\frac{1}{Q_{rt}^{\lambda}}i_t^{\iota},$$

where i_t is given in (B.2). Substituting for i_t in (B.2) we can also express the value function as a forward looking difference equation

$$\begin{aligned}\upsilon_{rt} &= \frac{1}{\sigma}\frac{1}{Q_{rt}}\frac{E_{rt}}{N_{rt}}+\left(\frac{\iota-1}{\iota}\right)w_{rt}^{R}\frac{1}{\zeta_r^I Z_t^I}\frac{1}{Q_{rt}^{\lambda}}\left(\frac{1}{1+r_t}\frac{\upsilon_{rt+1}}{w_{rt}^{R}}Q_{rt}^{\lambda}\zeta_r^I Z_t^I\right)^{\frac{\iota}{\iota-1}}\\ &= \frac{1}{\sigma}\frac{1}{Q_{rt}}\frac{E_{rt}}{N_{rt}}+\left(\frac{\iota-1}{\iota}\right)\left(\frac{Q_{rt}^{\lambda}\zeta_r^I Z_t^I}{w_{rt}^{R}}\right)^{\frac{1}{\zeta-1}}\left(\frac{\upsilon_{rt+1}}{1+r_t}\right)^{\frac{\iota}{\iota-1}}.\end{aligned}$$

B.1.2 Deriving the labor supply relationships

To derive the results in Section 2.4.2, we rely heavily on the max stability of the Frechet distribution. In particular, if the S dimensional vector $[x_s]_s$ is iid Frechet distributed across s, with

$$P\left(x_s \leq z\right) = F_s\left(z\right) = e^{-h_s z^{-\theta}},$$

the variable

$$a = \max_s \left[\lambda_s x_s\right] \tag{B.3}$$

is distributed according to

$$P\left(a \leq \alpha\right) = e^{-\Lambda^\theta \alpha^{-\theta}}$$

where

$$\Lambda = \quad \sum_s \lambda_s^\theta h_s \Bigg)^{1/\theta}$$

Using this result we can derive the expressions for average income, average human capital and the relative employment shares. In particular,

$$E\left[a\right] = \Gamma\left(1 - \frac{1}{\theta}\right) \times \Lambda$$

Deriving average income W_{nr}^k Total income of individual i is given by

$$y^i = \left\{w^P x^P, w^R x^R \quad .\right.$$

Hence,

$$P\left(y^i \leq \overline{y}\right) = e^{-W^\theta \overline{y}^{-\theta}}$$

where

$$W^i = \left(\left(w^P\right)^\theta h_i^P + \left(w^R\right)^\theta h_i^R\right)^{1/\theta}.$$

Hence, if for example individual i if skill type k and nationality n in region r, W^i is given by

$$W_{rn}^k = \left(\left(w_r^P\right)^\theta h_n^{Pk} + \left(w_r^R\right)^\theta h_n^{Rk}\right)^{1/\theta}.$$

Deriving the total supply of human capital by occupation (equation (2.15)) To derive the aggregate supply of human capital of individuals of type (n, k) in occupation j, note that the *average* number of efficiency units provided to occupation o is

$$E\left[z_n^{ok} | y_n^{jk} = \max_j \left\{y_n^{jk}\right\}\right] = E\left[z_n^{ok} | z_n^{ok} = \max_j \left\{\frac{w^j}{w^o} z_n^{jk}\right\}\right].$$

Using (B.3) with $\lambda_j = \frac{w^j}{w^o}$ this object is given by

$$\begin{aligned} E\left[z_n^{ok} | y_n^{jk} = \max_j \left\{y_n^{jk}\right\}\right] &= \Gamma\left(1 - \frac{1}{\theta}\right) \left. \sum_j \left(\frac{w^j}{w^o}\right)^\theta h_n^{jk}\right)^{1/\theta} \\ &= \Gamma\left(1 - \frac{1}{\theta}\right) \frac{\left(\sum_j \left(w^j\right)^\theta h_n^{jk}\right)^{1/\theta}}{w^o} \end{aligned}$$

Also note that the share of people of type (n, k) working in occupation j in region r is given by

$$s_{rn}^{jk} = P\left(w_r^j z_n^{jk} = \max_o \left\{w_r^o z_n^{ok}\right\}\right) = \frac{h_n^{jk} \left(w_r^j\right)^\theta}{\sum_j h_n^{jk} \left(w_r^j\right)^\theta} = h_n^{jk} \frac{\left(w_r^j\right)^\theta}{\left(W_{rn}^k\right)^\theta}.$$

Hence, letting the mass of workers of type (n, k) in region r working in occupation j be

$$L_{rn}^{jk} = L_{rn}^k s_{rn}^{jk},$$

we get that

$$\begin{aligned} H_{rn}^{jk} &= L_{rn}^{jk} E\left[z_n^{ok} | y_n^{jk} = \max_j \left\{y_n^{jk}\right\}\right] \\ &= L_{rn}^k s_{rn}^{jk} \Gamma\left(1 - \frac{1}{\theta}\right) \frac{W_{rn}^k}{w_r^j} \\ &= L_{rn}^k \Gamma\left(1 - \frac{1}{\theta}\right) s_{rn}^{jk} \left(\frac{h_n^{jk}}{s_{rn}^{jk}}\right)^{1/\theta} = L_{rn}^k \Gamma\left(1 - \frac{1}{\theta}\right) \left(h_n^{jk}\right)^{1/\theta} \left(s_{rn}^{jk}\right)^{\frac{\theta-1}{\theta}}. \end{aligned}$$

Alternatively we can also express this object as

$$H_{rn}^{jk} = L_{rn}^k \Gamma\left(1 - \frac{1}{\theta}\right) s_{rn}^{jk} \frac{W_{rn}^k}{w_r^j} = L_{rn}^k \Gamma\left(1 - \frac{1}{\theta}\right) h_n^{jk} \left(\frac{w_r^j}{W_{rn}^k}\right)^{\theta - 1}.$$

These expressions are exactly (2.15).

B.1.3 Proof of Lemma 2.4.3

The production function for the final good in (2.2) is given by

$$Y_{rt} = Z_{rt} X_{rt} = Z_{rt} \left(\int_{i=0}^1 x_{rt}(z)^{\frac{\sigma-1}{\sigma}} dF(z)\right)^{\frac{\sigma}{\sigma-1}}.$$

Letting U_{rt} be the price index for the bundle X_{rt}, we get that

$$\frac{u_{rt}(z)\, x_{rt}(z)}{U_{rt} X_{rt}} = \left(\frac{u_{rt}(z)}{P_{rt}}\right)^{1-\sigma} \Longrightarrow x_{rt}(z) = \left(\frac{u_{rt}(z)}{P_{rt}}\right)^{-\sigma} X_{rt}.$$

Substituting the production function (2.3) and $\frac{u_{rt}(z)}{P_{rt}} = \left(\frac{z^{\sigma-1}q}{Q}\right)^{\frac{1}{1-\sigma}}$, we get that total employment of production workers at firm i is given by

$$\begin{aligned} h_{rt}(z) &= z^{-1}q_{rt}(z)^{-\frac{1}{\sigma-1}}y(z) \\ &= z^{-1}q_{rt}(z)^{-\frac{1}{\sigma-1}}\left(\frac{z^{\sigma-1}q_{rt}(z)}{Q_r}\right)^{\frac{\sigma}{\sigma-1}}X_{rt} \\ &= z^{\sigma-1}q_{rt}(z)\left(\frac{1}{Q_r}\right)^{\frac{\sigma}{\sigma-1}}X_{rt} \end{aligned}$$

Hence, total labor demand is

$$\int h_{rt}(z)\,dF(z) = \left(\frac{1}{Q_{rt}}\right)^{\frac{\sigma}{\sigma-1}}X_{rt}\int z^{\sigma-1}q(z)\,dF(z) = \left(\frac{1}{Q_{rt}}\right)^{\frac{\sigma}{\sigma-1}}X_{rt}Q_{rt} = (Q_{rt})^{\frac{-1}{\sigma-1}}X_{rt}.$$

Labor market clearing implies that $\int h_{rt}(z)\,dF(z) = H_{rt}^P$. Hence,

$$X_{rt} = (Q_{rt})^{\frac{1}{\sigma-1}}H_{rt}^P.$$

Substituting into the production function yields

$$Y_{rt} = Z_{rt}(Q_{rt})^{\frac{1}{\sigma-1}}H_{rt}^P.$$

B.1.4 Characterization of the Balanced Growth Path (BGP)

In this section, we characterize the details of the balanced growth path (BGP). Along the BGP the allocation of people is constant across space and all aggregate variables grow at some constant rate, g^i where i is the relevant variable and i could be potentially different for different i. Along the balanced growth path interest rates are also constant, $r_t = r$. We assume that $\frac{Z_{rt+1}^A}{Z_{rt}^A} = 1 + \bar{g}_Z$, i.e. the exogenous component of productivity grows at rate g (which is the same for all regions). We also assume the aggregate research productivity M_t grows at a constant rate, i.e. $\frac{M_{t+1}}{M_t} = 1 + \overline{g}_M$

To have a balanced growth path we need that aggregate productivity A_{rt} grows at the same rate in all regions. Hence, see Lemma 2.4.3, we need that

$$1 + g_A = \frac{A_{rt+1}}{A_{rt}} = \frac{Z_{rt+1}^A}{Z_{rt}^A}\left(\frac{Q_{rt+1}}{Q_{rt}}\right)^{\frac{1}{\sigma-1}} = (1+\overline{g}_Z)(1+g_Q)^{\frac{1}{\sigma-1}}.$$

The BGP growth rate of productivity Q is therefore given by

$$1 + g_Q = \left(\frac{1+g_A}{1+g_Z}\right)^{\sigma-1}. \tag{B.4}$$

Using (2.23), this implies that the innovation rate in region r, i_{rt}, has to be constant across locations and time. Using (2.11) this implies that

$$1 + g_Q = i = \left(\frac{1}{1+r}\frac{\zeta_r^I Z_t^I}{Q_{rt}^{1-\lambda}}\frac{v_{rt+1}Q_{rt}}{w_{rt}^R}\right)^{\frac{1}{\iota-1}}. \tag{B.5}$$

Moreover, the value function v_{rt} given by equation (2.10) can be written as

$$\frac{v_{rt}Q_{rt}}{w^R_{rt}} = \frac{1}{\sigma}\frac{E_{rt}}{w^R_{rt}} + \frac{\iota-1}{\iota}\frac{Q^{1-\lambda}_{rt}}{\zeta^I_r Z^I_t} i^{\iota}. \tag{B.6}$$

First note that w^R_{rt} and E_{rt} grow at the same rate in a stationary equilibrium. Hence, conjecture that along a BGP both $\frac{v_{rt}Q_{rt}}{w^R_{rt}}$ and $\frac{Q^{1-\lambda}_{rt}}{Z^I_t}$ are constant. This implies that

$$1+g_Q = \left(1+\overline{g}_M\right)^{\frac{1}{1-\lambda}}. \tag{B.7}$$

(B.4) therefore implies that the growth rate of TFP is given by

$$1+g_A = \left(1+\overline{g}_Z\right)\left(1+\overline{g}_M\right)^{\frac{1}{1-\lambda}\frac{1}{\sigma-1}}. \tag{B.8}$$

Under our price normalization, wages w^R_{rt} and total spending E_{rt} are growing at the rate of aggregate TFP g_A. Hence, for $\frac{v_{rt}Q_{rt}}{w^R_{rt}}$ to be constant, the growth rate of the value function given by

$$\begin{aligned}
\frac{v_{rt+1}}{v_{rt}} &= \left(1+g_v\right) \\
&= \frac{w^R_{rt+1}/w^R_{rt}}{Q_{rt+1}/Q_{rt}} \\
&= \frac{1+g_A}{1+g_Q}\frac{v_{rt+1}}{v_{rt}} \\
&= \frac{\left(1+\overline{g}_Z\right)\left(1+\overline{g}_M\right)^{\frac{1}{1-\lambda}\frac{1}{\sigma-1}}}{\left(1+\overline{g}_M\right)^{\frac{1}{1-\lambda}}} \\
&= \left(1+\overline{g}_Z\right)\left(1+\overline{g}_M\right)^{\frac{1}{1-\lambda}\left(\frac{2-\sigma}{\sigma-1}\right)}
\end{aligned} \tag{B.9}$$

In fact, the value function can be solved explicitly along the BGP. Let $r > g_A$. Along the BGP the value function is given by

$$\frac{v_{rt}Q_{rt}}{w^R_{rt}} = \frac{1}{1-\frac{\zeta-1}{\zeta}\frac{1+g_A}{1+r}}\frac{1}{\sigma}\frac{E_{rt}}{w^R_{rt}}.$$

Proof. From (B.5) and the fact that v_{rt} grows at at rate $1+g_v$ given in (B.9), we get that

$$1+g_Q = i = \left(\frac{1}{1+r}\frac{\zeta^I_r Z^I_t}{Q^{1-\lambda}_{rt}}\frac{v_{rt+1}Q_{rt}}{w^R_{rt}}\right)^{\frac{1}{\iota-1}} = \left(\frac{1}{1+r}\frac{\zeta^I_r Z^I_t}{Q^{1-\lambda}_{rt}}\frac{\left(1+g_v\right)v_{rt}Q_{rt}}{w^R_{rt}}\right)^{\frac{1}{\zeta-1}}.$$

Hence, we can solve for $\frac{Q^{1-\lambda}_{rt}}{\varphi^I_r M_t}$ as

$$\frac{Q^{1-\lambda}_{rt}}{\zeta^I_r Z^I_t} = \frac{1}{1+r}\frac{1}{\left(1+g_Q\right)^{\zeta-1}}\frac{\left(1+g_v\right)v_{rt}Q_{rt}}{w^R_{rt}}.$$

Substituting into (B.6) yields

$$\begin{aligned}\frac{v_{rt}Q_{rt}}{w_{rt}^R} &= \frac{1}{\sigma}\frac{E_{rt}}{N_{rt}w_{rt}^R}+\frac{\iota-1}{\zeta}\frac{1}{1+r}\frac{1}{(1+g_Q)^{\zeta-1}}\frac{(1+g_v)\,v_{rt}Q_{rt}}{w_{rt}^R}i^{\zeta}\\ &= \frac{1}{\sigma}\frac{E_{rt}}{N_{rt}w_{rt}^R}+\frac{\iota-1}{\iota}\frac{1+g_Q}{1+r}\frac{(1+g_v)\,v_{rt}Q_{rt}}{w_{rt}^R}\end{aligned}$$

so that

$$\frac{v_{rt}Q_{rt}}{w_{rt}^R} = \frac{1}{1-\frac{\iota-1}{\iota}\frac{1+g_Q}{1+r}(1+g_v)}\frac{1}{\sigma}\frac{E_{rt}}{N_{rt}w_{rt}^R}=\frac{1}{1-\frac{\iota-1}{\iota}\frac{1+g_A}{1+r}}\frac{1}{\sigma}\frac{E_{rt}}{N_{rt}w_{rt}^R}.$$

□

The spatial productivity distribution Along the BGP, the spatial distribution of productivity Q_{rt} is stationary as all regions grow at the same rate. The *level* of productivity is, however, determined endogenously. In particular, (B.5) implies that

$$\frac{\varphi_r^I M_t}{Q_{rt}^{1-\lambda}}\frac{v_{rt+1}Q_{rt}}{w_{rt}^R}=\frac{\varphi_j^I M_t}{Q_{jt}^{1-\lambda}}\frac{v_{jt+1}Q_{jt}}{w_{jt}^R}\text{ for all } r,j. \tag{B.10}$$

This implies that relative productivities Q_{rt} and Q_{jt} are given by

$$\frac{Q_{rt}}{Q_{jt}}=\left(\frac{\varphi_r^I}{\varphi_j^I}\times\frac{\frac{v_{rt}Q_{rt}}{w_{rt}^R}}{\frac{v_{jt}Q_{jt}}{w_{jt}^R}}\right)^{\frac{1}{1-\lambda}}.$$

Hence, long-run differences in productivity across space Q_{rt}/Q_{jt} are governed by

$$\ln\left(\frac{Q_{rt}}{Q_{jt}}\right)=\underbrace{\frac{1}{1-\lambda}\ln\left(\frac{\varphi_r^I}{\varphi_j^I}\right)}_{\text{Exogenous differences in research efficiency}}+\underbrace{\frac{1}{1-\lambda}\ln\left(\frac{\frac{v_{rt}Q_{rt}}{w_{rt}^R}}{\frac{v_{jt}Q_{jt}}{w_{jt}^R}}\right)}_{\text{Endogenous value of innovation}}. \tag{B.11}$$

Using Proposition B.1.4, (B.11) can be written as

$$\ln\left(\frac{Q_{rt}}{Q_{jt}}\right)=\underbrace{\frac{1}{1-\lambda}\ln\left(\frac{\varphi_r^I}{\varphi_j^I}\right)}_{\text{Exogenous differences in research efficiency}}+\underbrace{\frac{1}{1-\lambda}\ln\left(\frac{\frac{E_{rt}}{N_{rt}w_{rt}^R}}{\frac{E_{jt}}{N_{jt}w_{jt}^R}}\right)}_{\text{Endogenous Market Size}}. \tag{B.12}$$

Finally, note that we can express (B.12) also in terms the labor supply. To do so note that along the BGP payments to researchers and production workers are equalized. Consider a BGP. Then

$$\frac{w_{rt}^R H_{rt}^R}{w_{rt}^P H_{rt}^P}=\frac{1}{\sigma-1}\frac{\frac{1}{\iota}\frac{1+g_A}{1+r}}{1-\frac{\iota-1}{\iota}\frac{1+g_A}{1+r}}.$$

Proof. The total demand for efficiency units in the research sector is given by

$$\frac{1}{\zeta_r^I Z_t^I}\frac{q}{Q_{rt}^{\lambda}}\frac{i_t^{\iota}}{\iota}w_{rt}^R$$

$$H_{rt}^R = \int_q \frac{1}{\zeta_r^I Z_t^I}\frac{q}{Q_{rt}^{\lambda}}\frac{i^{\iota}}{\iota}dF_{rt}(q) = \frac{Q_{rt}^{1-\lambda}}{\zeta_r^I Z_t^I}\frac{i^{\iota}}{\iota}.$$

Using that $i = 1 + g_Q$ and $\frac{Q_{rt}^{1-\lambda}}{\zeta_r^I Z_t^I} = \frac{1}{1+r}\frac{1}{\left(1+g_Q\right)^{\zeta-1}}\frac{(1+g_v)v_{rt}Q_{rt}}{w_{rt}^R}$ (see proof of Proposition B.1.4), we get that

$$H_{rt}^R = N_{rt}\frac{1+g_Q}{1+r}\frac{(1+g_v)\,v_{rt}Q_{rt}}{w_{rt}^R}\frac{1}{\iota} = \frac{1+g_A}{1+r}\frac{v_{rt}Q_{rt}}{w_{rt}^R}\frac{1}{\iota}.$$

Using the result in Proposition B.1.4, this implies that

$$H_{rt}^R w_{rt}^R = \frac{1+g_A}{1+r}\frac{1}{1-\frac{\iota-1}{\iota}\frac{1+g_A}{1+r}}\frac{1}{\sigma}\frac{1}{\iota}E_{rt}.$$

Hence, the payments to researchers are a constant fraction of revenue along the BGP. And because production workers also receive a constant fraction of revenue (see (2.19)), we get

$$\frac{H_{rt}^R w_{rt}^R}{w_{rt}^P H_{rt}^P} = \frac{1}{\sigma-1}\frac{\frac{1}{\iota}\frac{1+g_A}{1+r}}{1-\frac{\iota-1}{\iota}\frac{1+g_A}{1+r}}.$$

□

Using Proposition B.1.4, the long run distribution of productivity in (B.12) can also be expressed in terms of (endogenous) amount of resources in research

$$\ln\left(\frac{Q_{rt}}{Q_{jt}}\right) = \underbrace{\frac{1}{1-\lambda}\ln\left(\frac{\varphi_r^I}{\varphi_j^I}\right)}_{\text{Exogenous differences in research efficiency}} + \underbrace{\frac{1}{1-\lambda}\ln\left(\frac{H_{rt}^R}{H_{jt}^R}\right)}_{\text{Resources employed in research}}, \tag{B.13}$$

or the number of production workers and the relative cost of research

$$\ln\left(\frac{Q_{rt}}{Q_{jt}}\right) = \underbrace{\frac{1}{1-\lambda}\ln\left(\frac{\varphi_r^I}{\varphi_j^I}\right)}_{\text{Exogenous differences in research efficiency}} + \underbrace{\frac{1}{1-\lambda}\ln\left(\frac{H_{rt}^P}{H_{jt}^P}\right)}_{\text{Market size by production workers}} + \underbrace{\frac{1}{1-\lambda}\ln\left(\frac{w_{rt}^P/w_{rt}^R}{w_{jt}^P/w_{jt}^R}\right)}_{\text{Relative cost of research}} \tag{B.14}$$