UNIVERSITY OF CALIFORNIA SAN DIEGO

Essays in Urban and Regional Economics

A dissertation submitted in partial satisfaction of the
requirements for the degree
Doctor of Philosophy

in

Economics

by

Jiajun Lu

Committee in charge:

Professor Richard T. Carson, Chair
Professor Judson Boomhower
Professor Joshua Graff Zivin
Professor Ruixue Jia
Professor Isaac William Martin

2020

The dissertation of Jiajun Lu is approved, and it is acceptable in quality and form for publication on microfilm and electronically:

_____

_____

_____

_____

_____

Chair

University of California San Diego

2020

DEDICATION

To my parents for their everlasting love and support;

And to those who shared unforgettable moments in this journey.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGEMENTS

I would like to express my deepest gratitude to my main advisor, Richard T. Carson, for his guidance and support during my Ph.D. life. I am also sincerely grateful to Ruixue Jia, Judson Boomhower, and Julie Cullen for their support and recommendation on my job market. Landing a job would not be that smooth without their help. My other thanks go to the advisors in my committee, Josh Graff Zivin and Isaac William Martin, and anyone else who provided valuable comments and insightful advice on the development of the dissertation.

On a more personal level, I would also like to thank my fellow doctoral students in the Ph.D. program for their advice on the job market, helpful discussions on research, and warm friendship, including but not limited to Xiaxin Wang, Zhenting Sun, Wenbin Wu, Wei You, Ying Feng, Qihui Chen, Xueying Lu, Fanglin Sun, Xu Zhang, Chu Yu, Youngju Lee, and Xiao Ma. Moreover, I am grateful to my office mates, Linyan Zhu, Wonhyong Choi, and Nobuhiko Nakazawa, for providing such a great environment to maintain research productivity.

Chapter 1, in full, is currently being prepared for publication. Jiajun Lu. The dissertation author was the sole author of this chapter.

Chapter 2, in full, has been recently accepted for publication. Jiajun Lu. The dissertation author was the sole author of this chapter.

Chapter 3, in full, is a coauthored work with Richard T. Carson. Richard T. Carson; Jiajun Lu. It is currently in preparation for publication and the dissertation author was the primary researcher of this chapter.

VITA

2012	B.S. in Economics, University of Science and Technology Beijing
2014	M.A. in Finance, Zhejiang University
2020	Ph.D. in Economics, University of California San Diego

PUBLICATIONS

Jiajun Lu, "The Value of a South-Facing Orientation: A Hedonic Pricing Analysis of the Shanghai Housing Market", *Habitat International*, 81.1 (2018): 24-32.

ABSTRACT OF THE DISSERTATION

Essays in Urban and Regional Economics

by

Jiajun Lu

Doctor of Philosophy in Economics

University of California San Diego, 2020

Professor Richard T. Carson, Chair

Chapter 1 examines the economic consequences of expanding housing supply in productive urban cities and analyzes how residential sorting plays a role in forming a new market equilibrium. Using the newly released 2013-2017 American Community Survey data, I construct an economic model system that includes the models characterizing household residential location choices and their simultaneous spatial interactions with local labor markets, housing markets, and urban amenities across geographical areas in California. I find that, in an open economy with agglomeration effects, the positive residential sorting largely undoes what the housing legislation aims to achieve and reduces the quality of urban amenities in productive cities.

Chapter 2 documents the relationship between climate amenities and locational choices

in retirement. Using data from 2017 release of the American Community Survey, I construct a household residential location choice model and value climate amenities from the trade-offs among housing cost, climate amenities, and other locational attributes in a metropolitan statistical area (MSA). The results show that values of climate amenities vary with household demographic characteristics, and older households with a higher retirement income and disability have a higher marginal willingness to pay for a favorable climate. Using projected climate data, I find that over 2% of retired households would relocate in response to this level of climate change.

Chapter 3 investigates how the residential real estate market, the second-largest asset market, in the U.S. has been fundamentally changed by the advent of online real estate websites. Using data on over 50,000 completed transactions obtained from Zillow, we first look at how the availability of the Zestimate influences both listing and sales prices. The factors influencing the listing realtor's decision to hold an open house are examined, as is the role such an open house has on the sales price and sales timing. Empirical results suggest that Zestimates play an important and complex role in driving the sales process and that holding an initial open house substantially increases sales price and decreases time on the market.

Chapter 1

General Equilibrium Effects of an Urban Housing Supply Expansion: The Role of Residential Sorting

1.1 Introduction

Population migration has long been an engine of economic growth over the past century in the United States. The movement of the population seeking higher levels of well-being not only largely improved the social welfare but also formed a convergence in economic development across regions (Barro et al., 1991). However, this regional convergence has slowed considerably due to the decreasing population flows to high-income places over the past thirty years (Ganong and Shoag, 2017). The high housing and living expenses in most popular migration destinations result in a higher economic barrier that makes finding a chance at socioeconomic mobility in an affordable area much more difficult. As a state with the largest economy in the U.S.,[1] California has seen a diverging economic development across geographical areas. The increase

[1] https://en.wikipedia.org/wiki/List of U.S. states and territories by GDP.

in spatial wage dispersion is mainly driven by some coastal cities, like San Francisco and Los Angeles, which experienced strong growth in labor productivity (Moretti, 2012). These productive metropolitan areas provide many employment opportunities but also feature high growth rates of housing prices.[2] The high housing costs in these urban areas make local housing service the least affordable and the resulting affordability crisis becomes a growing concern from local residents, especially from the economically disadvantaged households.[3] It has long been under debate what causes high housing costs in urban areas, and many studies have sought to justify the high urban housing price. Some early works argue that higher housing prices capitalize better urban amenities in a spatial equilibrium (Rosen, 1979; Roback, 1982). Others propose that housing regulations and supply constraints can explain high urban rents and housing prices since many metropolitan areas adopted land-use restrictions that significantly constrained new housing supply (Glaeser et al., 2005; Glaeser and Ward, 2009; Furth and Gonzalez, 2019; Diamond et al., 2019). On top of that, some recent papers point out that, rather than the control oversupply of residential land, a large-scale positive assortment of high-quality workers causes the high housing costs in urban cities (Gyourko et al., 2013).

To get around the housing affordability crisis, the government of California recently proposed an ambitious bill, Senate Bill 50 (SB 50), in an attempt to reduce urban housing prices.[4] This proposed housing legislation primarily aims to override existing local zoning requirements in favor of new urban planning that allows for a higher residential density. It claims that the new housing policy can drive down high housing prices in metropolitan areas by substantially increasing the supply of housing. The rule of thumb that increasing housing supply reduces housing price has long been enshrined in folk wisdom. This idea from a simple mental model with

[2]Figure 1.22 in Appendix presents the long-lasting upward trends of household income and housing expense between metro and non-metro areas in California over the past decade.

[3]Figure 1.23 in Appendix shows the positive relationship between household income and housing cost across geographical areas in California. According to the National Association of Realtors (NAR), San Francisco in California has the lowest housing affordability index across the U.S. in 2016. https://www.nar.realtor/research-and-statistics.

[4]https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200SB50

a supply-demand equilibrium appears in many housing policies and has been in widespread use by Federal and state government agencies. However, these housing policies proposed typically by government officials or politicians oversimplify potential economic consequences of a housing supply expansion and thus become greatly at odds with the objective reality. Despite some appealing statements in it, the bill is likely to employ just a naïve cure for a very complicated real-world problem. Admittedly, in a closed city, building new housing units drives down housing prices, holding the local aggregate housing demand fixed. On the contrary, in an open-city system, households are mobile and can make locational decisions to improve the living environment. In response to a lower urban housing price, households are incentivized to move into the urban area and fill up as many dwelling units as newly built, as long as improved welfare outweighs moving costs. Numerous household relocations caused by the housing supply shock would, in turn, influence the demand side of the urban housing market. A large number of in-migrants increase the local population density and further strengthen urban agglomeration effects (Ganong and Shoag, 2017). These agglomeration externalities would be reflected in the wage premium due to enhanced productivity in the local labor market, which increases the purchasing power of housing services. Therefore, the potential large-scale population migration and urban agglomeration effects make it unclear whether more housing supplied can eventually drive down the housing price. Due to the agglomeration economy, the government's expansion policy can come into play the way exactly opposite to what it aims to do or make the housing prices even higher than before the expansion. To solve the puzzle in urban economics, this paper aims to answer the question of whether and why sometimes expanding housing supply can eventually drive down local housing prices in productive urban cities.

Apart from the uncertainty in its effectiveness, the new housing legislation that attempts to expand housing supply leaves out some unintended externalities on local residents, and a lot of statements made in this bill are likely to be misconceived. First, the massive residential sorting may largely reduce the quality of urban life. Due to a housing expansion, the inflow migration of

population into urban cities increases local residential density and substantially further worsens traffic congestion.[5] Moreover, a lower housing price increases the potential risk of large-scale defaults by new homebuyers as the values of their properties fall below what they owed mortgage lenders. These changes in household utility determinants and economic consequences that are not mentioned in the proposed housing legislation also influence household locational decisions, which would eventually be reflected in local housing prices. Therefore, the spatial interaction between population migration and local urban amenities further complicates the influence of a housing expansion on the local economy, from the perspective of general equilibrium. The complexity in the entire economic system comprised of multiple components motivates urban economists to explore their simultaneous and spatially interdependent relationships. It is of great importance to answer the question of what would happen to metropolitan cities if the housing supply were substantially expanded by building a sizable number of new dwelling units in the areas targeted. To this end, this paper comprehensively examines the economic consequences of an urban housing supply expansion in cities with a housing affordability crisis and analyzes how residential sorting plays a role in forming a new housing market equilibrium.

Many recent economic papers have sought to model the residential sorting and its complex interaction with a local economy (Jeanty et al., 2010). An important caveat to the relevant analysis is that people usually move for a variety of factors, and thus, asking them to choose a single reason could be misleading. Instead, in addition to the living environment of the residence, a household residential location determines a workplace, education resource, transport mode, and living expenditure, and so forth. These locational attributes become an inextricably entangled weave of mutual interdependencies and constraints, all depending on each other and with varying degrees of similarity and substitutability. Therefore, from the perspective of utility maximization, existing studies have provided evidence showing that household locational choices are simultaneously driven by employment opportunity (Head et al., 1995), commuting cost (So et al., 2001), housing

[5] Another negative externality associated with the higher population densities is the potential overcrowding of public facilities, like the beach and public parks.

service (Knapp et al., 2001), and neighborhood characteristics (Bayer et al., 2016). Despite the rapid development of research in modeling the residential location choice, relatively less attention has been paid to analyze how numerous household relocations form a population migration at an aggregate level and how it influences the housing market.[6] Only a few papers in recent years have analyzed the interaction between population migration and housing price dynamics (Jeanty et al., 2010; Mussa et al., 2017). On account of information network and spatial crowding effects, the impact of population displacement on housing prices can be highly spatially correlated across housing markets, and most empirical results are estimated under a spatial econometric framework.[7] However, these studies leave out the influence of an increased population density and resulting agglomeration externalities on the quality of urban amenities, which is another important determinant of local housing prices and further residential sorting.

No past research, to the best of my knowledge, has ever fully investigated household locational choices interacting with local housing markets, labor markets, and urban amenities across geographical areas. To fill the gap, I construct an economic model system connecting all four economic components to account for their simultaneous interactions. This is the first paper that proposes a theoretical framework and provides empirical evidence from a spatial economic system that coherently considers the interdependence among household locational decisions, local housing and labor markets, and urban amenities. Firstly, using the newly released data from 2013-2017 American Community Survey by the U.S. Census Bureau, I model household locational choices depending on a variety of factors that influence household utility if they were moving there.[8] It includes household income, housing price, potential moving cost, commute cost, and quality of local urban amenities. Modeling the choice behavior identifies the pattern of household residential location choices and demonstrates why expensive metro areas are still

[6]Most analyses mainly focus on the spatial effect of internal migration on local labor markets (Borjas, 2006; Ottaviano and Peri, 2012).

[7]Immigrants share the information of local locational attributes and thus influences the migration flow and neighboring housing markets.

[8]https://www.census.gov/programs-surveys/acs/

preferred for households in certain demographic groups. The empirical evidence, therefore, manifests how the consistent residential sorting is internally formed by numerous individual utility-maximizing locational decisions. Second, I specifically quantify the effect of population migration on the local labor market and traffic condition whereby agglomeration effects. Many studies have established that urban agglomerations yield to a significant wage premium (Wheaton and Lewis, 2002; Puga, 2010). To identify an agglomeration economy in urban areas, I estimate a wage equation on demographic factors and local population density. As part of the quality of urban amenities, the local traffic congestion index is calculated by the commute time regressed on the changing population density due to residential sorting. Lastly, I estimate a spatial housing price equation with respect to local demographic compositions, housing supply level, and local urban amenities, including traffic conditions.

Solving the economic models enables me to examine: 1) how households make trade-offs among various attributes when making a locational decision; 2) whether there exists a residential sorting in response to a shock in a local housing price; 3) why local demographic compositions are predictive of housing prices and household incomes going up in urban areas; 4) whether there exist agglomeration effects in a local economy; 5) how the traffic condition depends on the population density and impacts the housing price. Conditional on estimation results, I simulate the economic consequences of an expansion on urban housing supply in productive urban areas and explore the dynamic evolution of the economic indicators in the local economy. In response to the new price parameters in the housing markets, households would relocate to remaximize the household utility, which forms a large-scale residential sorting. As a result, the changing demographic compositions would reflect in local labor and housing markets, which, in return, influences the locational choices. The economic model system repeatedly solves itself until it comes to a new equilibrium. Then, I evaluate the new steady-state to see whether the housing supply expansion eventually reduces urban housing prices and examines its welfare impact.

Empirical results show that households prefer a residence with a higher household income,

lower housing price, lower cost of work commute, and higher-quality urban amenities, especially less traffic congestion. There exists a positive residential sorting where high-income, well-educated, and dual-worker households are sorted into metropolitan areas, while lower-income and less-educated households move out over time. A temporarily lower housing price drives a massive in-migration of highly productive households. Estimation results also confirm the existence of an urban wage premium due to an agglomeration economy. Moving more high-skilled households into these metropolitan areas would further raise their incomes and make local housing prices even higher than before the housing policy is deployed in productive urban areas. These findings suggest that, in an open-city economy with agglomeration effects, large-scale residential sorting can substantially undo what the housing legislation aims to do. In addition to its ineffectiveness in productive urban areas, an expansion in housing supply increases the local population density, resulting in a higher level of traffic congestion. This unintended consequence would further exacerbate local traffic conditions and cause a negative externality on existing residents. Lastly, the welfare gains of in-migrants from a housing supply expansion cannot offset the welfare loss of urban natives, while the entire population in California can benefit through numerous household relocations.

Under the coherent framework of an economic model system, this paper characterizes household residential location choices and their interaction with a local economy. It contributes to the existing literature in the following ways. Firstly, the economic model system constructed in the paper addresses the simultaneous nature of household residential location choices, housing and labor markets, and the quality of urban amenities, which is an important advantage over existing methods in measuring local housing prices. Secondly, using the mixed logit model, a state-of-the-art modeling method that accounts for preference heterogeneity, I explore heterogeneous housing preferences across demographically identifiable groups and provide a micro-foundation for taste-based sorting driven by household utility-maximizing locational choices. Third, this paper models household residential location choices at a geographical scale in California finer

than any other estimations, using newly released census data with locations identified at the level of PUMA within metropolitan areas.[9] Other studies have analyzed residential sorting at the level of metro areas (Knapp et al., 2001; Sinha et al., 2017) and states (Davies et al., 2001) in the U.S. market, which surely underestimates the number of households moving across local labor and housing market boundaries within a metropolitan area.[10] Therefore, this paper largely improves the modeling precision by exploring the residential sorting at such disaggregated level in California. At last, this paper presents many new estimates on the willingness to pay (WTP) for locational attributes and other urban amenities at smaller geographic scales.

The remainder of this paper is organized as follows. Section 2 presents a brief review of relevant literature. The empirical background and theoretical framework are introduced in section 3 and 4, respectively. Section 5 describes the data used for model estimations. Section 6 details the model specifications and estimation procedure. Empirical results are presented in section 7. In section 8, I simulate the economic consequences of a housing supply expansion at a moderate level and investigate the residential sorting in response to a lower urban housing price. The paper concludes with a summary of key findings and policy implications.

1.2 Literature Review

This section presents an extensive review of three strands of economic literature related to this paper, including household residential location choice, the economic consequences of population migration, and the model system of urban spatial economics.

[9] Public Use Microdata Area (PUMA) is a statistical geographic area defined in the American Community Survey that contains at least 100,000 people. It is measured at the spatial level lower than in metropolitan areas. Figure 1.24 in Appendix shows the geographical boundaries of PUMAs in the San Diego metropolitan area and its comparison with San Diego County.

[10] Over 70 percent of migrants have moved within a metropolitan area, while around 20 percent of migrants would even have a short-distance move within a smaller area, as shown in Figure 1.3.

1.2.1 Household Residential Location Choice

Modeling the household residential location choice has long been the focal point in analyzing residential sorting. The residential location choice model regularly characterizes the selection of residential location by weighting site attributes of each available alternative for economically rational consumers (McFadden, 1978). Previous studies have mainly focused on what motivates and affects the household moving decision and locational choice, and a large number of determinants with predictive power on migration have been examined. Existing researches have found that residential location choices can be motivated primarily by employment opportunity (Greenwood et al., 1991), education resources (Benabou, 1993), retirement (Duncombe et al., 2001), and other neighborhood characteristics (Rabe and Taylor, 2010; Bayer et al., 2016). These empirical researches typically incorporate local housing prices as one of the locational attributes, assuming a homogeneous housing preference across households. However, many find that there exist diverse preferences on housing service and preference heterogeneity in housing plays a critical role in the locational choice (Abraham and Hunt, 1997; Goodman and Thibodeau, 1998). In addition to affordable housing, some other studies focus on the influence of work commute (So et al., 2001) and workplace (Yates and Mackay, 2006; Frenkel and Kaplan, 2015) on home location choices.

With the development of econometric techniques, there are essentially two methodologies that are widely used in estimating household preference parameters (Hensher and Johnson, 2018). The first approach is to assume that there exist spatial equilibrium in both housing and labor markets and analyze trade-off among alternatives in a hedonic pricing setting (Rosen, 1974). The second approach, a discrete choice model, can be estimated without having to assume that a housing market is in equilibrium (Anas, 1983). Much research has found that, due to large transaction costs and search friction, housing markets are actually in disequilibrium but are always moving towards a new equilibrium through residential sorting (Hill and Syed, 2016). Therefore, repeatedly observed household relocations across geographical areas justify the use of a discrete

choice model in a setting of market disequilibrium.

Over the past decades, the analysis of household residential location choices has been largely facilitated by the advancement of discrete choice modeling methods (Hensher et al., 2005). Since the seminal paper by McFadden (1973), the conditional multinomial logit (MNL) model has been the most common approach to modeling home location choice, due to its computational advantage. The MNL model imposes the assumption of independence of irrelevant alternatives (IIA) among alternatives, making cross-elasticity across each pair of choice alternatives equivalent. However, the IIA assumption is often violated if households perceive some destination alternatives as closer substitutes (Daly and Zachary, 1978). To address the issue of IIA constraint, several more advanced discrete choice models were developed, such as ordered generalized extreme value (OGEV) model (Small, 1987), cross-nested logit (CNL) (Vovsha, 1997), and paired combinatorial logit (PCL) model (Wen and Koppelman, 2001). Nevertheless, none of them account for potential heterogeneous preferences for attributes. To accommodate preference heterogeneity, McFadden and Train (2000) propose a mixed logit model that incorporates random preference coefficients. This flexible model features a framework that captures an unrestricted substitution pattern and individual-specific preferences in the decision-making process. Due to its appealing property, the mixed logit model has been widely adopted in modeling a discrete choice in various contexts, including recreational activity (Bhat and Gossen, 2004), electricity supplier choice (Revelt and Train, 2000), and driving behavior (Behnood et al., 2016). However, few papers have sought to model household location choices with preference heterogeneity in site attributes using this state-of-the-art modeling method.

The development of the choice modeling approach facilitates the analysis of microdata and contributes to a substantial body of literature on household locational choice, covering a wide range of residential areas. Existing researches have provided relevant empirical evidence outside of the U.S., such as Canada (Liaw and Ledent, 1987), Australia (Yates and Mackay, 2006; Ho and Hensher, 2014) and Israel (Frenkel and Kaplan, 2015). Many studies also analyzed locational

choice in the context of U.S. housing market, including some that explore residential sorting in a local area, such as Texas (Zhou and Kockelman, 2008) and Florida (Rapaport, 1997), and others working on interregional migration across metro areas (Knapp et al., 2001; Sinha et al., 2017) and states (Davies et al., 2001) in the entire United States.

1.2.2 Economic Consequences of Population Migration

The second strand of economic literature related to this paper is mainly focused on the interaction between the local housing market and population migration formed by numerous residential location choices. Many previous papers first explore what drives large-scale internal migration. A large body of econometric evidence suggests that amenities of high quality along with high earnings in a local labor market attract migrants while relatively high housing prices repel them (Clark and Hunter, 1992; Cameron et al., 2006). On the other hand, many recent papers have sought to analyze the influence of residential sorting and population migration. Boustan et al. (2010) study the effect of internal migration on the U.S. labor market, showing that migration had little effect on the average income of existing residents and crowded some residents out of the local labor market. As another important economic outcome, the local housing cost can also be influenced by the massive relocations, and the last few years have seen an increased focus on its influence on the housing market. Molloy et al. (2011) summarize the recent studies on the trend of internal migration in the U.S. and other countries and discuss how relocation activities interact with both labor and housing markets. Some papers quantify the influence of migration on the U.S. housing market by constructing a general equilibrium model of population migration and housing price dynamics, showing that the inflow of immigrants raises the housing costs for local residents (Jeanty et al., 2010; Ottaviano et al., 2012). Other than the U.S. housing market, Chen et al. (2011) explore the role of migration on the housing prices in China's urbanization and find that the internal population displacement significantly enlarges the gap in the housing prices across geographical regions in China. Akbari and Aydede (2012) assess the impact of

immigration on the housing market in Canada and find a significant but small positive influence on prices of privately-owned dwellings.

Many previous studies ignored the spatial interactions between migration and the housing market. DeSilva et al. (2012) propose that leaving out spatial dependence in house prices across neighborhoods can bias the estimated influence of internal population migration. Along with the development of spatial econometrics, recent years have seen some researches that perform a similar analysis under the framework of spatial econometrics, accounting for spatially correlated local housing prices. Among spatial econometric techniques, the spatial error model (SEM) and the spatial autoregressive model (SAR) become two popular methods, each containing one type of spatial interaction effect (Elhorst, 2010). A spatial autoregressive disturbance is incorporated in SEM, while the SAR controls for spatial dependence by adding a spatially lagged dependent variable as an additional explanatory variable. Soon after, many spatial econometrics studies have shifted to a spatial Durbin model (SDM) that includes both spatially lagged price variables and attribute variables (LeSage and Pace, 2009).[11] The SDM, composed of the SAR and SEM, accommodates a more flexible spatial dependence pattern and thus enables the researchers to analyze direct, indirect, and total marginal effects of housing price determinants. By estimating the SDM, some studies have explored the determinants of a local housing price, such as labor market accessibility in Norway (Osland and Thorsen, 2013) and air pollution in Spain (Fernández-Avilés et al., 2012). More recent work by Mussa et al. (2017) exploits both the direct and indirect effects of migration flows on the U.S. housing market at the level of metro areas, pointing to the importance of demographic variations in the migration population for a local housing price.

1.2.3 Economic System Model

The third group of relevant literature is the development of an economic model system. A local area includes many elements, such as households, labor market, transportation, and real

[11] Anselin (1988) first proposed the spatial Durbin model.

estate development. Since these economic components are highly interdependent, many studies have attempted to construct an integrated economic system that jointly models these components. The development of its work dates back to Orcutt (1957), who proposed an original framework of the economic system. The baseline model was then improved by Alonso (1960) and extended to a conceptual model of a large-scale metro area by Lowry (1964). These early urban system models are essentially descriptive economic models that cannot be used to evaluate an urban policy quantitatively. With the advancement of modeling techniques and computing capacity, many quantitative models used for empirical analyses have been later developed, attempting to incorporate explicit representations of an extensive range of elements in the entire economic system.

Based on the urban structure in the U.S., Putman and Ducca (1978) first developed the DRAM model integrated with a multinomial logit model to calculate the location surplus measure. This DRAM model concentrates on household home location choices in connection with transportation and land use in an urban area. Rosen (1979) and Roback (1982) later constructed a Rosen-Roback model that characterizes the connections among prices of amenities, wages, and prices across cities. Following the travel demand function proposed by Ben-Akiva et al. (1985), Putman (1991) improved the urban framework in an employment allocation (EMPAL) model by adding the economic component of trip distribution and mode choice formulated as a nested logit model, in addition to home and employment location choices. Given the development of separate model systems, Putman and Chan (2001) later constructed a new model system, METROPOLIS, that contains DRAM, EMPAL, and other programs, including calibration procedures. This economic model system, embedded in a GIS environment, achieves a dramatic reduction in the difficulty of model application and enables urban planners to evaluate alternative transportation plans and policies on urban development and travel patterns (Krishnamurthy and Kockelman, 2003).

When assessing the consequences of urban policy on local development, municipal

planning agencies need models to be behaviorally clear and as transparent as possible. However, due to the absence of micro-foundations at the household level, most urban systems with a skeptical "black-box" model could not be easily explained to policymakers or the public (Lee, 1994). Household activities are always the main driving force of local economic growth and thus should be taken as the critical element of an economic model system (Pagliara et al., 2010). In an attempt to make urban models easier to interpret, many economic systems focus on micro-level household activities.

Given the availability of micro-level data of households, Waddell (2000) develops UrbanSim, a model system based on microsimulation.[12] It contains multi-modal modeling for housing choice, home location decision, and commute mode choice at the household level, using a range of specifications and econometric techniques. This utility-consistent economic system has recently been widely used in analyzing household behaviors and their connection with the urban development in land use (Schirmer et al., 2015), transportation (Jin and Lee, 2018), and environmental planning (Wang and Yuan, 2018). In a metropolitan-scale or even larger environment, household preferences and urban locational attributes vary widely across geographical areas. Thus, modeling interdependence among economic components needs to be localized. As a disaggregate model system, UrbanSim has a significant advantage over other economic model systems in being easily customized (Waddell et al., 2007).

1.3 Empirical Background

This section introduces the empirical background of this study, including the housing affordability issue and housing bill proposed by the State of California, the geography of the study area, and some stylized facts in residential sorting.

[12] https://www.urbansim.org

1.3.1 Housing Affordability Crisis

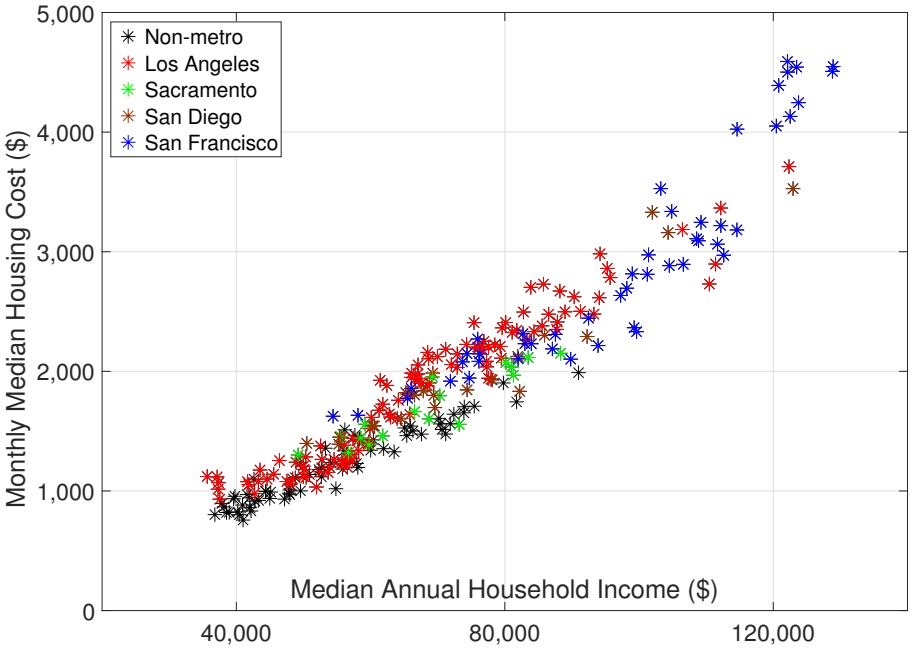
The past decades have seen fast economic growth in California, especially in urbanized regions. The high-income household living in these large metropolitan areas, however, usually have disproportionately high housing expenses. Figure 1.1 presents the relationships among household income, housing expense, and share of income on housing across geographical areas.[13] Panel (a) shows there exists a positive relationship between median household income and housing expense across PUMAs.[14] A majority of cities in metropolitan areas (colored) feature both higher income and housing cost than non-metro areas (black). Some cities in the greater San Francisco Bay Area have both the highest median household income and housing expense.[15] Given the same median household income, most non-metro areas have a lower median housing expense than metro areas. Panel (b) presents the relationship between median household income and the share of household income spent on housing service. It can be seen that, in general, households allocate a higher portion of expenditure to housing services in the high-income areas. Some highly productive urban areas, like San Francisco and Los Angeles, have the least affordable housing service, even if they are inhabited by many high-income households. On the contrary, lower-income households living in non-metro areas can afford more housing services.

High housing expenses in these productive metropolitan areas form an economic barrier that makes it much more difficult for households from outside to move in, which dramatically reduces the chance at socioeconomic mobility. In some highly productive cities, existing households who have relatively less income can afford the least housing service. The lack of affordable housing becomes a massive burden for many local economically disadvantaged households in

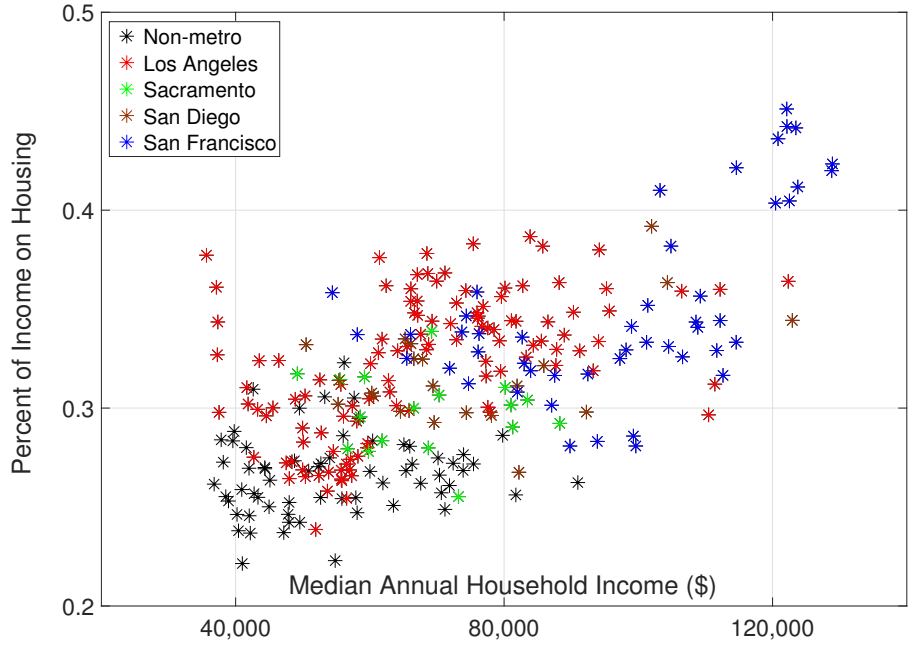
[13]The geographical unit, PUMA, is the lowest level of identifiable location in the American Community Survey. Figure 1.2 shows its geographical boundaries, which is introduced in detail later.

[14]Housing expense depends on a housing tenure choice. Housing cost paid by renters includes the rent, insurance, and utility fees, while homeowners pay property tax, homeowner association fee, insurance, utility fees, and, if applicable, a home mortgage.

[15]The San Francisco metropolitan area involves San Francisco, San Jose, and Oakland. The Los Angeles metropolitan area includes Los Angeles, Long Beach, and Riverside. San Diego, Carlsbad, and San Marcos constitute the San Diego metropolitan area. The Sacramento metropolitan area comprises Sacramento, Arden-Arcade, and Roseville.



(a) Household Income and Housing Cost



(b) Share of Household Income Spent on Housing

Figure 1.1: Relationship among Household Income, Housing Cost, and Share of Income on Housing Service

these urban areas, leading to a growing housing affordability crisis in California.

1.3.2 California's Housing Bill

In an attempt to address the shortage of affordable housing, California legislature recently proposed a Senate Bill 50 (SB 50) that promises a solution to California's high cost for housing services.[16] The high-profile Senate Bill 50, in essence, aims to drive down the urban housing prices by substantially expanding the housing supply there. This proposed aggressive policy comes with a bold promise to combat rising housing price by creating 3.5 million new housing units by 2025, which accounts for nearly 30% of the existing housing stock.[17] It forces the productive cities to relax their local zoning restrictions on increasing residential density in areas deemed job or transportation rich. Specifically, this proposed bill enables real estate developers to build multi-story residential properties near major transit stops or employment centers. It also allows the construction of more small apartments and townhouses, rather than single-family houses, near the job-rich neighborhoods. This housing bill got much public attention since it was proposed and has been introduced by multiple influential media, including Los Angeles Times[18] and VOX[19].

1.3.3 Geography of the Study Area

To explore the economic consequences of a housing supply expansion as proposed by Senate Bill 50, this paper takes California as the main study area. Figure 1.2 shows its geography where the entire state is partitioned into multiple non-overlapping Public Use Microdata Areas

[16] Scott Wiener, the California State Senator, introduced housing bill in December 2018 and amended it multiple times in early 2019.

[17] As of January 1, 2019, there are a total of 12,214,549 housing units, according to statistics given by the Department of Finance in the State of California. http://www.dof.ca.gov/Forecasting/Demographics/Estimates/E-5/

[18] https://www.latimes.com/politics/la-pol-ca-senate-bill-50-changes-20190424-story.html

[19] https://www.vox.com/policy-and-politics/2018/12/7/18125644/scott-wiener-sb-50-california-housing

(PUMAs). PUMA is a statistical geographic area that contains at least 100,000 people.[20] The metropolitan areas are shown in color, while the white areas represent non-metro regions. There exist a total of 265 PUMAs in California, and some of them constitute San Francisco, Los Angeles, San Diego, and Sacramento metropolitan areas. The standard delineations of metropolitan areas for the U.S. are designed by the United States Department of Agriculture. It identifies the metro-nonmetro status of each Public Use Microdata Area (PUMA).[21] The geographical size of a PUMA decreases as population density rises. In metropolitan areas, the geographical area of a county is comprised of multiple PUMAs and thus becomes too large to analyze the intra-metro residential sorting.[22] Figure 1.24 in Appendix shows the comparison between San Diego County and PUMAs in the San Diego metropolitan area. There are a total of 22 PUMAs in San Diego County, and there exist significant variations in housing expenses across PUMAs within the county. Analyzing migrations at the county level underestimates the number of people that move across local labor and housing market boundaries within a metropolitan area. Thus, taking PUMA, rather than the county, as geographic units can largely improve the accuracy of modeling household location choices.

1.3.4 Stylized Facts of Population Migration

In an open economic system in California, households are highly mobile, and population migrations take place in multiple ranges. Figure 1.3 shows the migration pattern within California in recent years. Panel (a) presents the proportions of the moving population, showing that about 19% of households living in California have moved in each year. Panel (b) presents the percents of movement in each range among the moving population. It is shown that around 70% of movements occur within a metropolitan area, while around 20% of relocations take place across

[20] The geographical boundaries of PUMAs are delineated by the U.S. Census Bureau based on the results of the 2010 Census. See: https://www.census.gov/geographies/reference-maps/2010/geo/2010-pumas.html

[21] https://www.ers.usda.gov/topics/rural-economy-population/rural-classifications

[22] The maximum number of PUMAs per county is 69 for Los Angeles County.

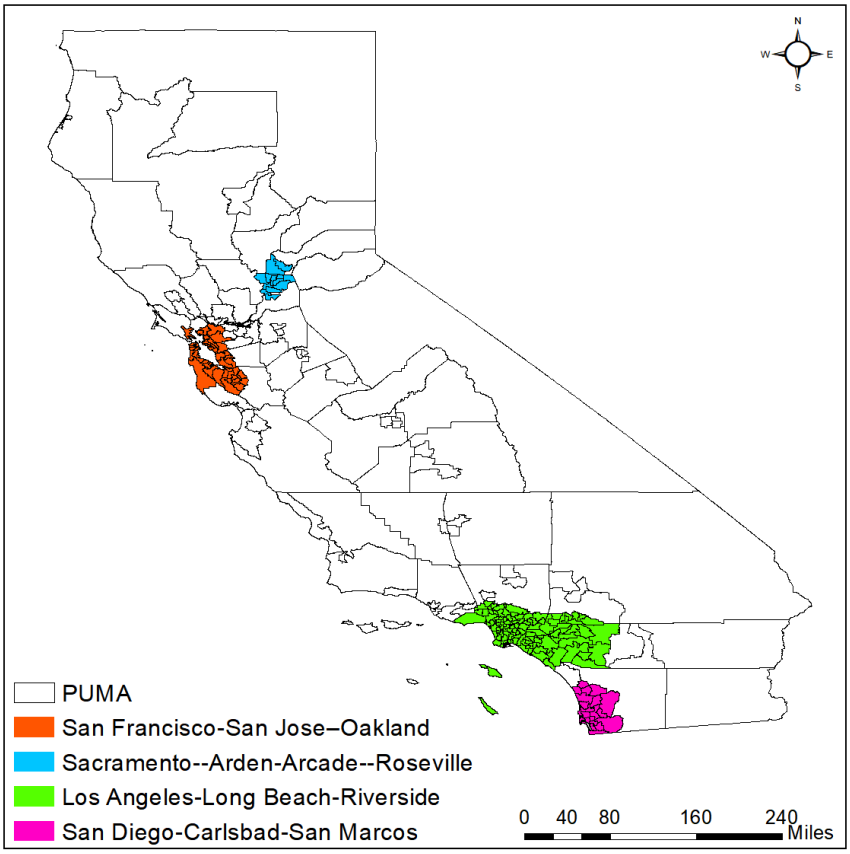
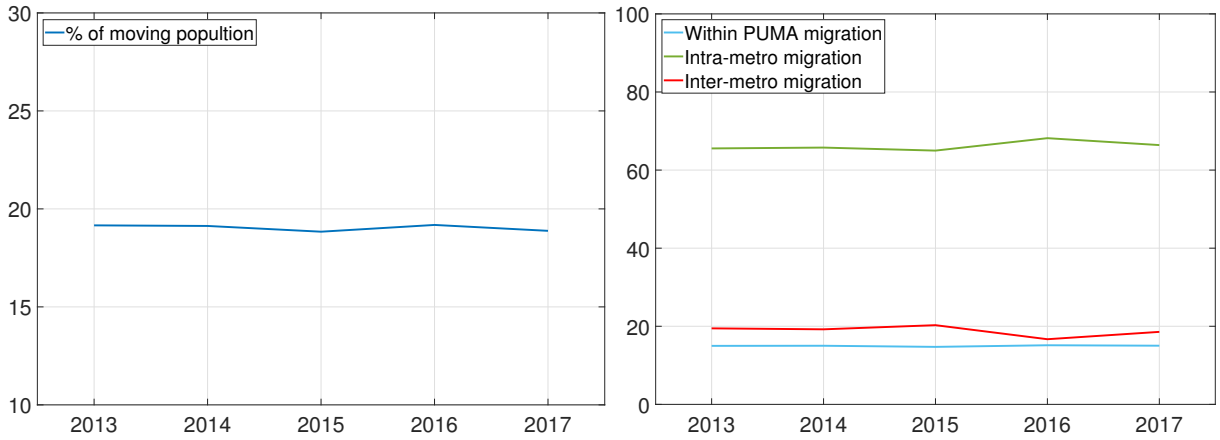


Figure 1.2: Geographical Boundaries of PUMAs in California

metropolitan areas. About 18% of migrants move in a shorter distance within a PUMA.



(a) Mobility Rate in California (b) Percent of Moving in Each Range

Figure 1.3: Migration Pattern within California

In reality, population migration can take place in a broader range, and households who choose someplace in California as the destination can move from anywhere in the U.S. or the rest of the world. For the completeness of potential migrations, origins include all states in the continental U.S., some U.S. islands, and the remaining areas in the world. Figure 1.4 shows the percents of households from each origin among the population moving into the San Diego metropolitan area during 2013-2017. It is shown that, over the past five years, about 94% of households who moved into the San Diego metropolitan area are from areas in California, where over 50% came from Los Angeles metropolitan area. Moreover, around 3.4% of households who moved to San Diego are domestic migrants, while 2.6% of them are international immigrants.

Among all households moving into the San Diego metropolitan area, there exist large variations in the probability of choosing each PUMA as the destination. Panel (a) in Figure 1.5 presents the percents of households outside San Diego moving to each PUMA in San Diego. It shows that, generally, the place with a higher monthly income, net of housing expense, attract more immigrants.[23] Panel (b) in Figure 1.5 shows the pattern of internal migrations within the

[23]The monthly net income equals the monthly median household income of existing residents minus the monthly median housing expense.

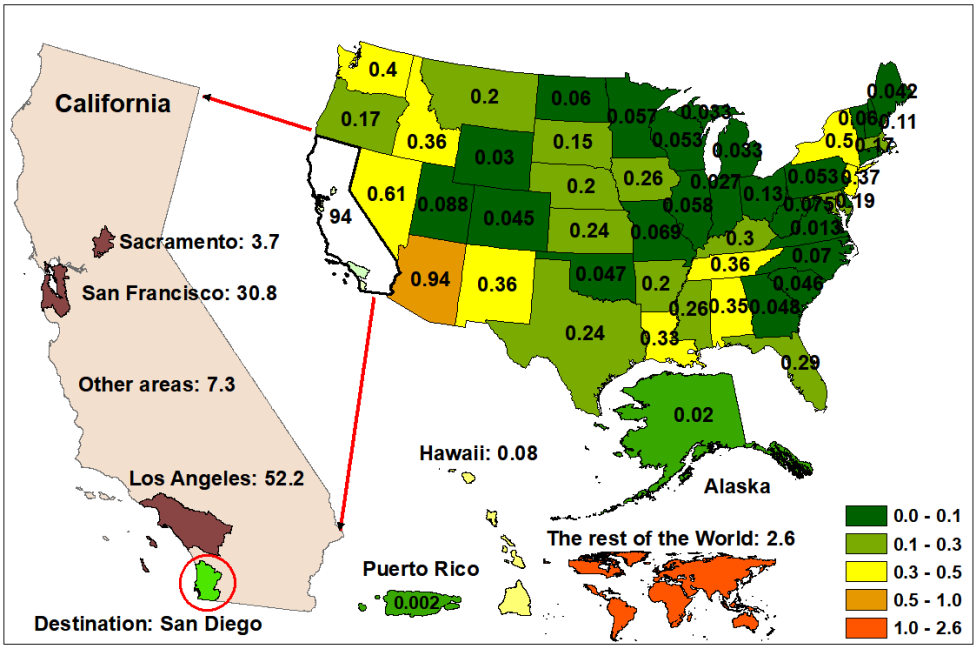
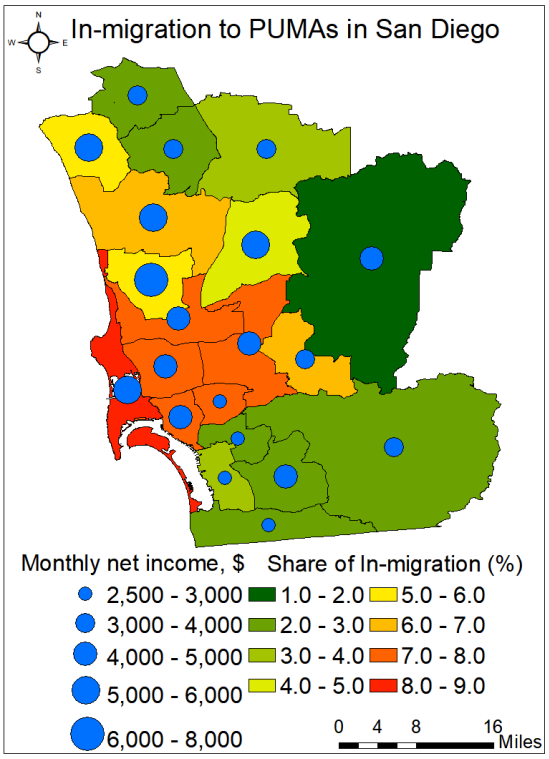


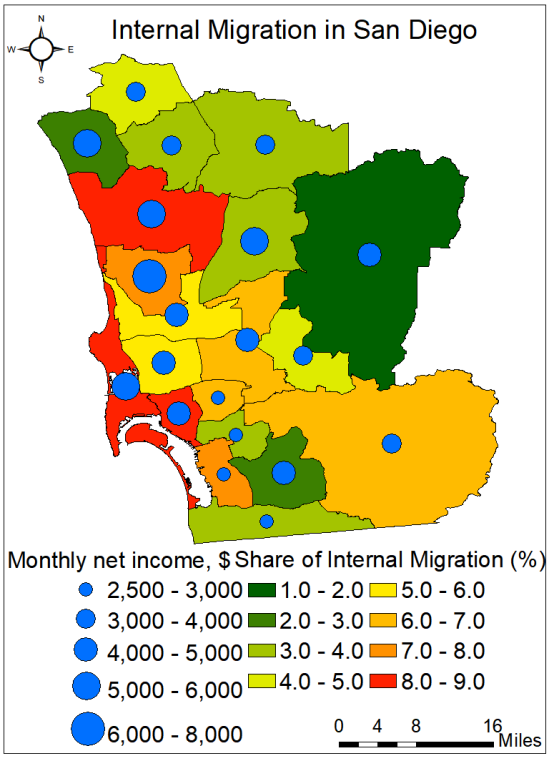
Figure 1.4: In-migration to San Diego Metropolitan Area

San Diego metropolitan area. Among households moving within the metropolitan area, there exist significant geographical variations in the percents of receiving the migrants. The existing residents in the San Diego metropolitan area have heterogeneous preferences over local communities when making a locational decision.

Table 1.1 shows the pattern of inter-metro migration in California. It presents, out of a certain metropolitan area, the percents of households who move to each destination. Among the population moving out of San Diego metropolitan area, over 75% of households move to Los Angeles metropolitan area, while almost 90% of households who move out of San Francisco select Los Angeles and Sacramento in total. This migration pattern between metropolitan areas in California implies that the moving distance between the origin and destination plays an important role in locational decisions.



(a) Share of In-migration (%)



(b) Internal Migration in San Diego

Figure 1.5: Migration Pattern in San Diego Metropolitan Area

Table 1.1: Moving between Metropolitan Areas in California

Out \ In	LA	SA	SD	SF	Total
LA		11.00%	44.68%	44.32%	100%
SA	41.60%		7.44%	50.96%	100%
SD	75.41%	5.01%		19.58%	100%
SF	52.79%	36.81%	10.40%		100%

Notes: The numbers represent, out of a certain metropolitan area, the percents of households who move to each destination. The abbreviations of metropolitan areas are as follows: LA (Los Angeles, Long Beach, and Riverside), SA (Sacramento, Arden-Arcade, and Roseville), SD (San Diego, Carlsbad, and San Marcos), and SF (San Francisco, San Jose, and Oakland). See Figure 1.2 for the geography.

1.3.5 Residential Sorting Pattern

Many studies have analyzed the effects of population migration on local housing and labor markets (Jeanty et al., 2010; Ottaviano et al., 2012). To explain why housing prices in metropolitan areas keep rising, I explore the pattern of residential sorting across geographical areas in California and demographic differences between households moving to metro areas and those moving to nonmetro areas. Figure 1.6 shows the percents of households who choose each PUMA among the moving population, along with a local annual median household income and monthly housing expense over the time period 2013-2017. The dots are colored based on the proportions of households who move to each PUMA. It is seen that more households prefer high-income areas, and these popular destinations feature both a high median household income and a high housing expense. This residential sorting pattern indicates that a higher housing price in an urban city seems to be mainly driven by a higher household income and housing demand by a larger population moving there.

Figure 1.6: Residential Sorting Pattern across PUMAs in California

Existing research has found that there exists positive residential sorting in multiple ways,

such as skills (Grogger and Hanson, 2011) and labor force participation (Johnson, 2014). To further seek the migration pattern in California, I decompose the entire moving population into households who move to metropolitan and nonmetro areas and examine the variations in household demographics between the two migration groups.[24] Table 1.2 shows that there exist significant differences in educational attainment and labor force participation between households moving to metro and nonmetro areas.[25] It can be seen that, as opposed to nonmetro areas, metropolitan areas have been gradually inhabited by well-educated households. The percent of householders who have a bachelor's degree is significantly higher in the population that moves to metropolitan areas than nonmetro areas. The difference in the percent of graduate degree holders is even larger between the two groups. In local labor markets, the households moving to metropolitan areas have a higher labor force participation rate than those moving to nonmetro areas.[26] Moreover, those who settle in metropolitan areas are slightly younger than nonmetro areas, even if the difference in age is barely significant.

Table 1.2: Metro-nonmetro Demographic Differences among Moving Population

Demographics	Metro	Nonmetro	Mean Difference[1]	All[2]
% of Bachelor's degree	41.66	34.94	6.72***	39.01
% of Graduate degree	18.13	5.84	12.29***	13.87
Number of earners	1.42	1.02	0.32***	1.29
Age of the householder	39.50	40.41	-0.91*	40.20

Note: Statistics are calculated over the moving population in 2013-2017, and economic variables are measured in 2017 U.S. dollars. [1]The significances of t-test for equality of means are reported, *** p<0.01, ** p<0.05, * p<0.1. [2]All represents the entire moving population.

In addition to the demographic differences between migrants into the metro and nonmetro areas, there could exist a consistent pattern of population displacement across localities in California. Figure 1.7 presents the demographic differences between in- and out-migrants in each

[24]Households are considered to be moving to a metropolitan area if they end up living in a metropolitan area, including those moving within metropolitan areas and from nonmetro areas.

[25]The educational attainment refers to the education background of a householder.

[26]The labor force participation is observed in the destinations. Admittedly, a household might have a different participation choice in another labor market, but the changes are not very common.

PUMA over 2013-2017.[27] Panel (a) and Panel (b) present the differences in the percent of a householder who has a bachelor's degree and graduate degree or higher, respectively, between in- and out-migrants in each PUMA. It is shown that households moving into high-income areas usually have a higher level of educational achievement than those who move out. The difference is more salient for the graduate degree or higher, which implies an education-based positive sorting pattern across geographical areas in California. Panel (c) shows that in-migrants to expensive areas tend to have more earners per household, as opposed to out-migrants.[28] Panel (d) illustrates that households moving into high-income areas are younger than those moving out, though the difference is relatively small. The demographic differences conform to the pattern of a positive sorting where well-educated and hard-working households are consistently sorted into high-wage and high-rent cities (Diamond, 2016).

1.4 Structure of the Underlying Economic Models

This section introduces the economic models developed to analyze the economic consequences of expansion in urban housing supply, with household residential location choices taken as a critical role in the general equilibrium framework.

1.4.1 Theoretical Framework

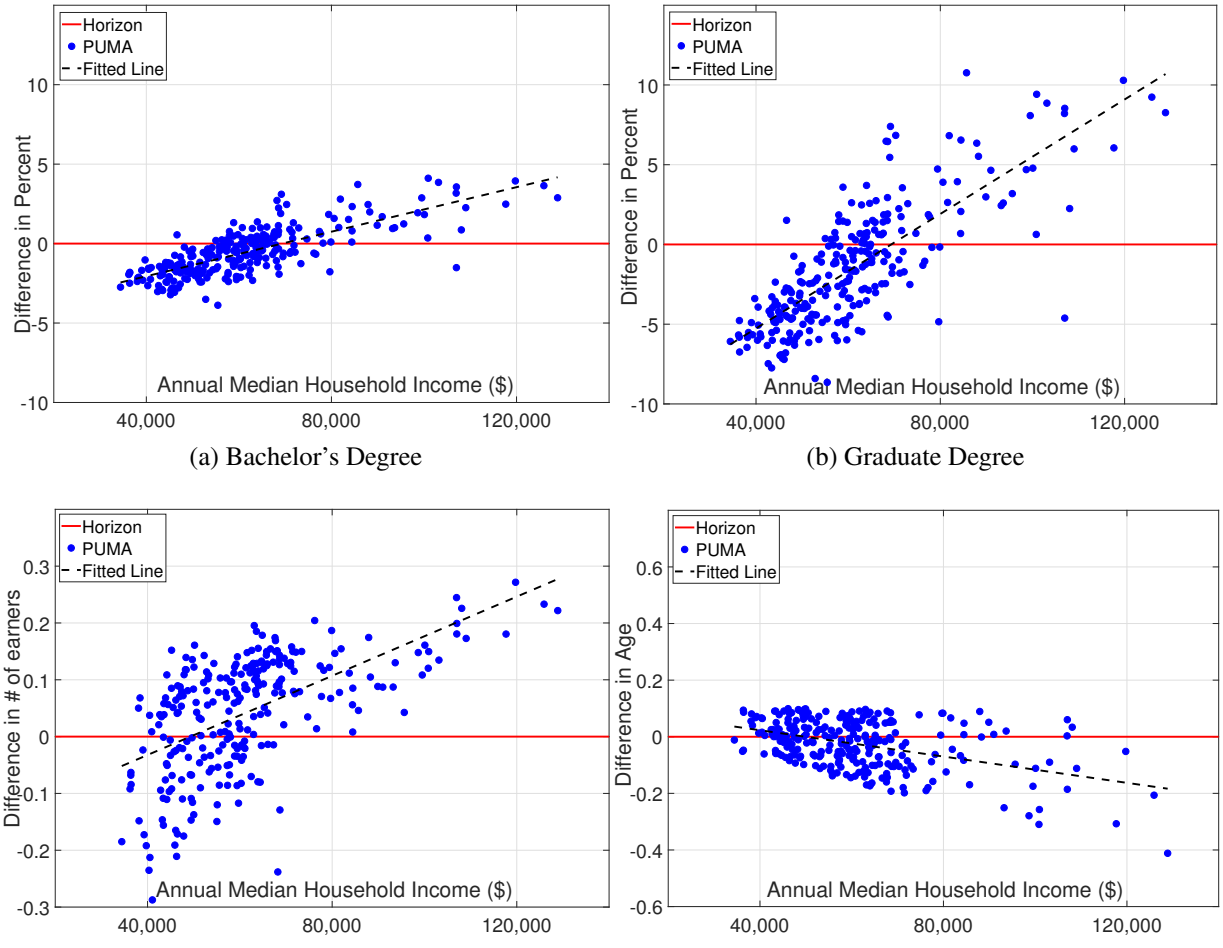
To characterize the structure of an open spatial economy, I construct an economic model system that integrates household locational choices, local housing and labor markets, and urban

[27] The differences in household demographic factors between in- and out-migrants in each PUMA are calculated as follows:

$$\Delta X_i = \bar{X}_i(\text{in}) - \bar{X}_i(\text{out}) = \frac{1}{M}\sum_{h=1}^{M} X_{hi} - \frac{1}{N}\sum_{h=1}^{N} X_{hi}, \tag{1.1}$$

where $\bar{X}_i(\text{in})$ and $\bar{X}_i(\text{out})$ denote the averages of household demographics among the population moving in and out of PUMA i. M and N are the numbers of households moving in and out of PUMA i. Data comes from the 2013-2017 American Community Survey.

[28] Labor force participation of both in- and out-migrants is measured in destinations, assuming households have the same labor participation in two places.



(a) Bachelor's Degree

(b) Graduate Degree

(c) Number of earners

(d) Age of the householder

Figure 1.7: Demographic Differences between In- and Out-migrants in Each PUMA

amenities across geographical areas. Figure 1.8 presents its theoretical conceptualization and the simultaneous interactions among the four economic components.

Firstly, I construct a household locational choice model, assuming that a locational decision depends on a variety of factors that influence household utility. It mainly includes household income, housing expenses, and the quality of local urban amenities. The numerous locational decisions made by moving households form massive population migration and thus change the local population density and demographic compositions. I assume that any single household in the housing market is too small to influence the local housing expense. But, in the aggregate, they can influence the housing expense. Second, I estimate the effects of population displacement on the local labor market, housing market, and traffic conditions, respectively. In the labor market, I estimate a wage equation where the household alternative income in a specific location, if they were moving there, is mainly determined by household demographics, local job structure, and a temporal trend. In addition, to account for an agglomeration economy in urban areas, I incorporate the population density of workers into the wage equation. In the housing market, I estimate a housing price equation with respect to local demographic compositions, housing supply levels, and locational attributes. Lastly, the local traffic condition is influenced by population density, local commuting system, demographic compositions. In return, the traffic condition also determines the work-to-home commute time and thus has an influence on household locational decisions. As a locational attribute, the level of traffic congestion also impacts the quality of urban amenities, which is eventually reflected in the local housing price.

1.4.2 Key Assumptions

The real world, specifically the situation in California, is more complex than the framework of the economic model system illustrated above. To simplify the economic system, several assumptions are made when characterizing the economic components.

The first key assumption is that the entire economic system is open with perfect mobility

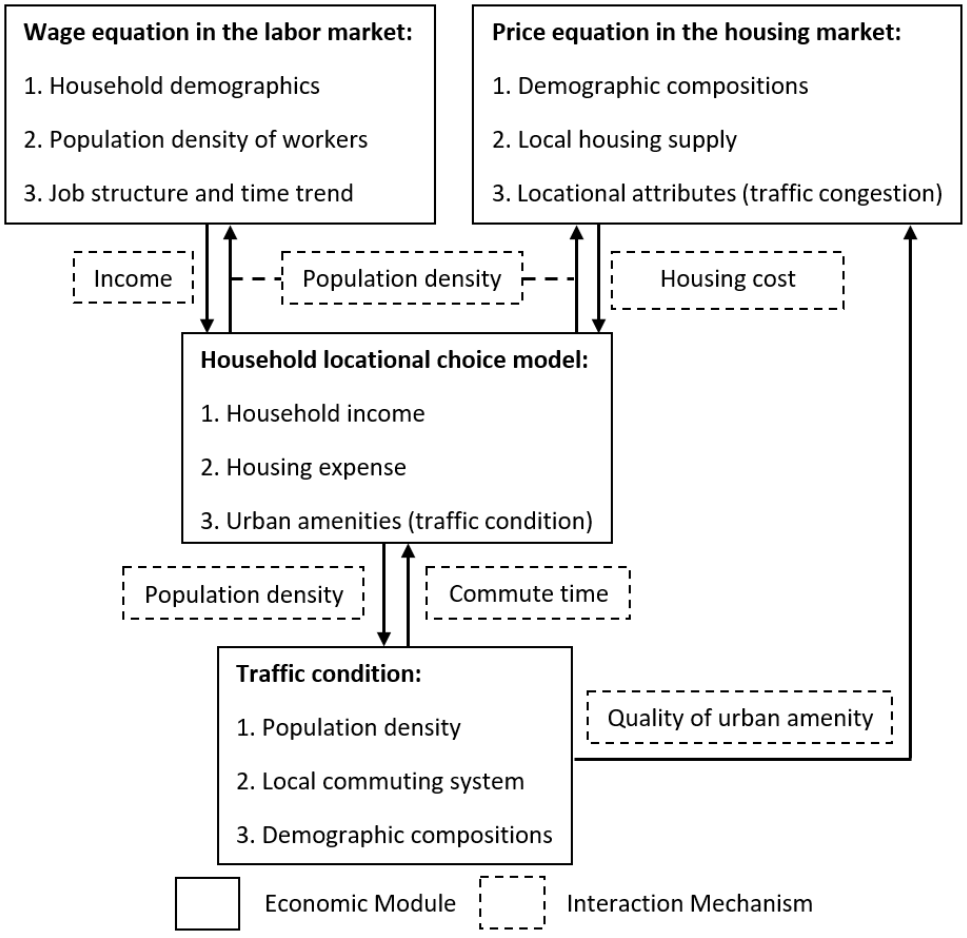


Figure 1.8: Economic Model System

across geographical areas in California.[29] In the setting, households are fully mobile and choose whichever city that maximizes their utility. For the completeness of the model, I assume that households moving from anywhere in the U.S. or the rest of the world into California will relocate within California, given that they have already decided to live in California.

Secondly, I assume that there exists an urban agglomeration effect in the labor market that impacts wage premium, and its mechanism does not change. This indicates that firms and workers in cities are productive and that there are benefits to being close to each other.[30] The positive externality of higher productivity resulting from residential sorting would be reflected in the household incomes and local housing prices.

The third assumption relates to household preferences. It is supposed that households have stable preferences for income, housing, and locational attributes over time when making locational choices. The preference parameters and willingness-to-pay functions are estimated from the representation of a household utility function. Moreover, working households are assumed to have homogeneous preferences for income and urban amenities, while there exists heterogeneity in housing preferences across households.

At last, I assume that an urban housing supply expansion caused by Senate Bill 50 is exogenous. When simulating the economic consequences of the housing policy, the economic system model takes an increase in the local housing supply level as a one-time shock. In reality, a city government can get the fund for some affordable housing program from tax revenue and substantially expand the housing supply by building many new housing units.

[29] It suggests the absence of the household registration system and households, in theory, can move to any place in California.

[30] To further understand what is happening, note that the same programmer is worth substantially more to an employer in the San Francisco Bay area than the same employer in Portland. The standard economic theory shows that, in a competitive market, the effective wages workers are paid equal to their marginal production, i.e., the additional net profit the programmer earns for the firm. The higher salary in the Bay Area comes from the agglomeration effect that results in the same programmer adding more to the company's bottom line located in its Bay Area office rather than its Portland office.

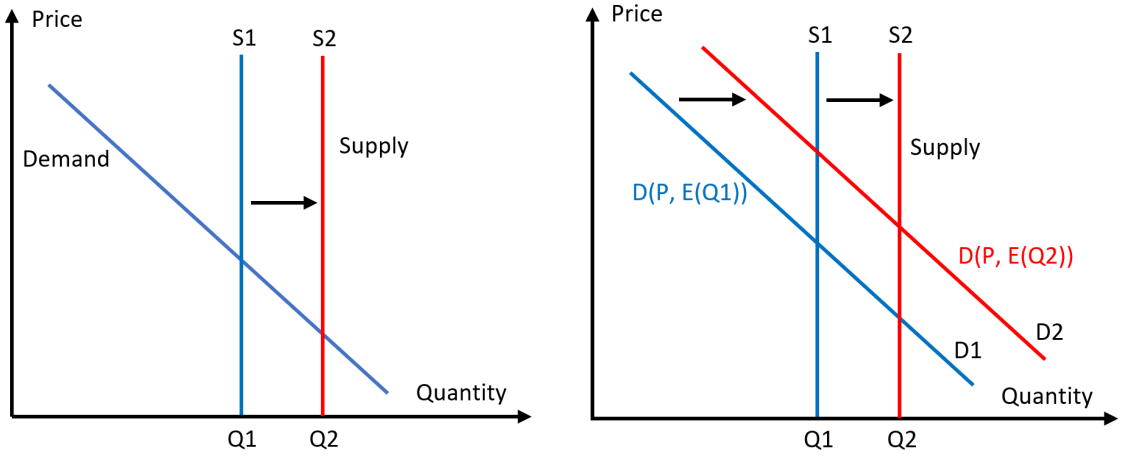
1.4.3 An Illustrative Example of Housing Price Dynamics

Based on the assumptions in the economic model system, an illustrative example is made to describe the dynamics of a local housing price according to the proposed housing bill. For simplicity, assume that the supply of local housing service is price inelastic and that the one-time housing expansion due to the housing policy is exogenous. The aggregate housing demand, $D(P,E(Q))$, depends on both a housing price, P, and an externality, E, that is a function of the city agglomeration, Q. Figure 1.9 illustrates the housing price dynamics with an expansion in housing supply and externality on housing demand. Panel (a) presents that the expansion of housing supply from the level $S1$ to $S2$ in stage 1 and that the housing quantity in the market equilibrium goes up from $Q1$ to $Q2$. In stage 2, the larger equilibrium quantity of housing, $Q2$, in the local economy enlarges the city agglomeration effects, which influences the aggregate housing demand in multiple ways, including a positive shock in the labor productivity and a negative impact on the local traffic condition.[31] These agglomeration effects result in the net externality, $E(Q2)$, and form a new housing demand curve, $D(P,E(Q2))$. Eventually, the new aggregate demand and supply of housing service lead to a new market equilibrium with an undetermined housing price movement. In an urban area, the net positive externality on the wage premium can substantially increase the aggregate housing demand, which is likely to make the new housing price even higher than before the housing supply expansion.

1.5 Data

This section introduces the data used to estimate the economic system model and characterize the spatial economy in California.

[31] Another example of a positive agglomeration effect on the local aggregate housing demand is the sports stadium, a public good that benefits more households.



(a) Housing Supply Expansion (b) Agglomeration Externalities on Demand

Figure 1.9: Housing Price Dynamics with a Housing Supply Expansion and Agglomeration Externality on Housing Demand

1.5.1 American Community Survey

The primary source of data used to construct the household location choice model and estimate household preference parameters is the Public Use Microdata Sample (PUMS) from the American Community Survey (ACS). It is a very detailed and comprehensive survey that describes household socioeconomic and demographic characteristics, covering around 1% of the entire population in the United States at household and personal level on a yearly basis.[32] This census data also contains rich information on household choices of workplace, residence, housing service, and commute time. I use the newly released data over the period 2013-2017. It contains information on 737,503 households and 1,872,509 residents in California.

1.5.2 Non-housing Expense

In addition to housing expenses, a household needs to pay for various non-housing services, which depends on both a locational choice and household size. I obtain the data of livable non-housing expense comes from Living Wage Calculator, a database that reports all

[32]https://www.census.gov/programs-surveys/acs/

expenses for a decent life.[33] It involves all basic needs and includes food, childcare, health care, transportation, other necessities, and income taxes. The expenses vary by household size and are measured at the county level. To make living costs compatible with locational choices, they are converted into population-weighted non-housing expenses at the PUMA level.

1.5.3 Moving and Commute Distance

When making a relocation decision to maximize household utility, the potential moving cost deters a long-distance movement and influences a household residential location choice. It needs to be controlled for in the model, although the data of actual moving costs are unavailable. Due to the data limitation, this paper uses a moving distance to proxy for the moving cost. Due to the highly developed highway system, most households pay a moving company or drive a vehicle to move their stuff in California. The economic cost of moving, mainly gasoline paid and time spent during the relocation, is proportional to the moving distance, which makes the model estimation still unbiased.[34] Moreover, when choosing a residence, a household also considers the daily commuting cost from home to work. Thus, the distance between the dwelling and workplace needs to be calculated.

The variable of distances is largely refined as follows. First, I calculate population-weighted centroids as geographic coordinates and take them as origins and destinations in measuring the distances.[35] This is to account for large geographical variations in population density within each PUMA. Then, instead of Euclidean distance between two geographic coordinates, I obtain the shortest driving distances to proxy for the actual moving distances more accurately. The shortest driving distances between two coordinates are obtained from the Distance Matrix API in Google Map Platform.[36]

[33] See: http://livingwage.mit.edu/

[34] Technically, the coefficients estimated with distances in a discrete choice model are the same as actual economic moving costs since the likelihood ratios are determined by relative numbers of variables for two alternatives.

[35] https://www.census.gov/geo/reference/centersofpop.html

[36] https://cloud.google.com/maps-platform/routes/?apis=routes

1.5.4 Locational Attributes

In addition to the census data, locational attributes are attained from multiple sources and controlled for in estimating household location choices and housing price equation. I include as many location-specific urban amenities that households care about as possible in the model. When estimating the housing price equation, I also control for numerous critical locational attributes that influence local housing expenses. All the variables are averaged over 2013-2017.

From the U.S. Census, I obtain the data of population density and housing supply density in each PUMA in California. The housing supply density represents the number of housing units supplied per square mile. I use the number of elementary schools to proxy for the quality of local education system and obtain the location information of schools from California Department of Education.[37] There are a total of 5,887 elementary schools in the 2017-2018 school year in California. The data concerning the number of public parks are provided by the California Department of Parks and Recreation.[38] I also calculate an index of the traffic congestion level using the data of commuting time.[39]

In addition to urban amenities, climate amenities also influence household utility and locational decisions. The climate data comes from GHCN-Daily, a dataset that contains rich climatic information provided by the NOAA National Climatic Data Center of the United States.[40] This paper adopts some common climatic attributes, including average temperature in winter and summer, precipitation, and snowfall. The mean temperatures in winter are measured over the three months from December to February, while the mean temperatures in summer are the average from June to August. All temperatures are measured in degrees Celsius. They are all included as utility determinants controlling for influences of these climate amenities.

[37] https://www.cde.ca.gov/ds/sd/cb/ceffingertipfacts.asp
[38] https://www.parks.ca.gov/
[39] The index of traffic congestion level is calculated in the equation (1.9), which is introduced later.
[40] https://docs.opendata.aws/noaa-ghcn-pds/readme.html

1.5.5 Descriptive Statistics

Table 1.3 reports the summary statistics and data sources of these variables. There are a total of 1,872,509 residents in 737,503 households surveyed during 2013-2017, living in one of the 265 PUMAs in California. Among the surveyed population, there are 942,411 employees in the labor market. It is seen that there are large variations in household income and expenses on housing and non-housing services across households. Housing costs are the total rental costs for renters, while homeowners report their rental equivalent estimates based on quality-adjusted new rental contracts.[41] On average, around 1.3 members take part in the labor force in a household. Most workers have a high school degree, and over a third of them are college graduates. A majority of respondents surveyed have U.S. citizenship and come from white households.

Regarding the locational choices, it shows that about 11% of households move out of a PUMA, and 4% move to a different metropolitan area. A household, on average, relocates to a residence 41 miles away from where they lived a year ago. For the work-home commute, the average time a worker spends on a one-way trip is about 27 minutes, and the average commute distance is nearly 14 miles.

1.6 Model Specification and Estimation Procedure

Following the economic model system in Figure 1.8, I detail the model specifications of these economic components and describe the estimation procedure.

1.6.1 Household Residential Locational Choice Model

In an attempt to explore the mechanisms of the consistent residential sorting, I first model a household residential location choice and provide the micro-foundations to explain why moving

[41] The expenditure a household allocates to the housing service is dependent on the housing tenure choice. Renters pay the rent, insurance, and utility fees, while homeowners pay property tax, homeowner association fee, insurance, utility fees, and, if applicable, a home mortgage.

Table 1.3: Summary Statistics of the Variables

Variable	Description	N	Mean	SD	Min	Max	Source
Household demographics							
Income	Household income[1]	737,503	71,729	55,114	2,000	1,900,000	ACS
Hcost	Housing cost	737,503	28,589	14,037	4,800	400,000	ACS
NonHcost	Non-housing cost	737,503	25,073	8,083	2,939	105,000	ACS
hsize	# of household members	737,503	2.65	1.41	1	7	ACS
age	Age of a householder	737,503	36.53	20.14	16	94	ACS
nwork	# of earners	737,503	1.29	0.61	0	2	ACS
female	Female	1,872,509	0.51	0.49	0	1	ACS
cit	Citizenship	1,872,509	0.87	0.34	0	1	ACS
white	White race	1,872,509	0.61	0.48	0	1	ACS
High	High school degree	942,411[2]	0.82	0.30	0	1	ACS
CollegeD	Bachelor's degree	942,411	0.39	0.48	0	1	ACS
GradD	Graduate degree	942,411	0.12	0.28	0	1	ACS
disability	With a disability	942,411	0.07	0.47	0	1	ACS
Locational choice and commute							
I^{PUMA}	Move out of a PUMA	737,503	0.11	0.13	0	1	ACS
I^{Metro}	Move out of a metro area	737,503	0.04	0.02	0	1	ACS
MovDis	Moving distance (mile)[3]	737,503	40.89	76.73	0	785	Google
ComTime	Commute time (minute)[4]	942,411	26.76	20.92	1	178	ACS
ComDis	Commute distance (mile)	942,411	13.90	15.74	0	70	Google
Locational attributes							
pden	Population density (p/mi^2)	265	246.98	154.61	7.46	14,316.02	Census
hden	Housing supply density[5]	265	74.03	55.74	3.41	7,384.91	Census
DisCoast	Distance to the coast (mile)	265	41.61	29.22	0.09	182.70	Google
educ	# of elementary schools	265	22.22	25.32	0	68	Govt[6]
park	# of parks	265	1.15	1.31	0	8	Govt
Traffic	Traffic congestion level[7]	265	0.55	0.24	0.00	1.00	Google
STemp	Summer temperature (°C)	265	22.13	3.31	12.11	38.42	NOAA
WTemp	Winter temperature (°C)	265	9.29	6.67	3.27	15.92	NOAA
precip	Annual precipitation (inch)	265	19.12	14.81	4.52	30.12	NOAA
snow	Annual snowfall (inch)	265	18.9	18.44	0.00	54.11	NOAA

Note: There are a total of 1,872,509 residents in 737,503 households surveyed during 2013-2017. [1]It is current household income measured in 2017 U.S. dollars. [2]Among the surveyed population, there are 942,411 employees in the labor market. [3]All the distance-related variables are obtained from the Distance Matrix API in Google Map Platform. [4]It is a one-way trip from home to work. [5]The housing supply density represents the number of housing units supplied per square mile. [6]It implies that the data is provided by an associated Department of California government. [7]The index of traffic congestion level is calculated in the equation (1.9), which is introduced later.

to metro areas has long been a utility-maximizing response for households. Following the seminal work by McFadden (1973), households are utility maximizers who attain utility through the selection of a preferred PUMA in California.

Assume that a household i, facing all available alternative locations of residence and workplace, chooses the home location j ($j \in \mathbf{J}$) and workplace k, if working, from all accessible places nearby ($k \in \mathbf{K}$) to obtain a certain level of utility U_{ijk}, if and only if this alternative yields the highest utility, i.e., $U_{ijk} > U_{ipq}$, $\forall\ j \neq p$ and $k \neq q$. Figure 1.10 illustrates the two-dimensional choice simultaneously made by a working household.[42]

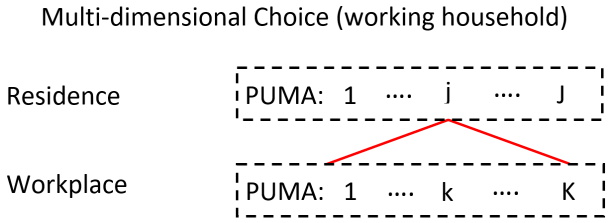


Figure 1.10: Joint Choice of Workplace and Residence by a Working Household

U_{ijk} is a stochastic variable, decomposed as the sum of a systematic component V_{ijk} and random part ε_{ijk}, i.e., $U_{ijk} = V_{ijk} + \varepsilon_{ijk}$. The systematic component, V_{ijk}, is a function of all observable attributes of alternatives and household characteristics, which is typically expressed in an additively separable and linear-in-parameters form. The random part, ε_{ijk}, captures heterogeneity in preferences that are unobserved. This paper assumes that a household utility is dependent upon household income, housing expense, expense on non-housing services, commute cost if working, moving cost in relocation if moving, and all locational amenities of the chosen residence. Specifically, the utility that household i receives when living in PUMA j and working in PUMA k is given by:

$$U_{ijk} = V_{ijk} + \varepsilon_{ijk} = (\text{Income}_{ik} - \text{NonHcost}_{ij})\alpha + \text{Hcost}_{ij}\beta_1 + \Gamma_{ijk} + M_{ij} + \eta_j + \varepsilon_{ijk}, \qquad (1.2)$$

[42]For retired households, they are assumed to disregard the workplace and only select a residence, as shown in Figure 1.26 in the Appendix.

where Income_{ik} is the expected permanent income household i can receive if working in PUMA k.[43] NonHcost_{ij} denotes the expense on all non-housing services. $\text{Income}_{ik} - \text{NonHcost}_{ij}$ is the Hicksian bundle that measures the disposable income or budget on housing service.[44] Hcost_{ij} represents the housing expense the household i pays if living in PUMA j. For simplicity, I assume that a household consumes the same bundle of housing services and take the average housing expense for the same housing choice to predict an alternative housing expenditure if moving to a different place.[45] Figure 1.25 in Appendix presents the summary statistics of housing choices. Γ_{ijk} is the generalized commute cost between residence j and workplace k, and $\Gamma_{ijk} = \text{ComDis}_{jk}\beta_2 + \text{ComTime}_{ijk}\beta_3$, where ComDis_{jk} and ComTime_{ijk} are the total commute distance and commute time in a one-way trip for household i.[46] Going forward, The moving cost of a relocation, M_{ij}, involves an economic moving cost and psychic cost of moving out of a place.[47] $M_{ij} = \text{MovDis}_{ij}\beta_4 + I_{ij}^{\text{PUMA}}\beta_5 + I_{ij}^{\text{Metro}}\beta_6$, where $\text{MovDis}_{ij}\beta_4$ is the economic moving cost that is proxied by a moving distance g_{ij}. $I_{ij}^{\text{PUMA}}\beta_5 + I_{ij}^{\text{Metro}}\beta_6$ represent the psychic cost, where I_{ij}^{PUMA} and I_{ij}^{Metro} denote dummy variables that equal one if a household moves out a PUMA and a metropolitan area. η_j is the locational fixed effect at PUMA level that controls for all location-specific attributes, including urban and climate amenities. ε_{ijk} is the error term that incorporates unobserved utility-related preference heterogeneity. This model leaves out time

[43] If there are multiple earners in a household, the household income becomes the sum of wages made by all earners, i.e., $\text{Income}_{ik} = \sum_w \text{Income}_{iwk} + b_i$, where Income_{iwk} is the wage income of the worker w. b_i is the non-wage income or retirement income of household i that is assumed to be unrelated to the location.

[44] For simplicity, I impose the constraint of equal coefficients on household income and the expense on non-housing service (Sinha et al., 2017).

[45] Admittedly, households may have different housing choices across cities. To test the validity of this assumption, I estimate the average number and standard deviations of some critical housing characteristics, e.g., the number of bedrooms and household tenure choice, across PUMAs for different demographic groups but find no significant variations. Specifically, I use one sample t test for each housing characteristic and fail to reject the null hypothesis of the same housing choice for each demographic group. Moreover, according to the American Housing Survey, there exist small variations ($\leq 5\%$) in housing adjustment across metropolitan areas. See: https://www.census.gov/programs-surveys/ahs.html

[46] For multi-worker households, the generalized commute cost becomes the sum of costs for all workers, i.e., $\Gamma_{ijk} = \sum_w \text{ComDis}_{jkw}\beta_2 + \text{ComTime}_{ijkw}\beta_3$, where ComTime_{ijkw} is the commute time of worker w in household i between residence j and workplace k.

[47] The psychic cost, which describes non-monetary costs, such as the loss of social network and familiarity within surrounding environment in the previous location, is largely dependent upon the range of movement (Davies et al., 2001).

fixed effects since they would cancel out in each discrete choice. Instead, the influence of time trend has been incorporated in estimating income and price variables.

When making a locational choice, a household typically puts considerable weight on the local housing expense. Many papers have provided evidence that households have heterogeneous preferences for housing (Abraham and Hunt, 1997; Goodman and Thibodeau, 1998). Based on the two stylized facts shown in Figure 1.1, in high-cost areas, households gain higher incomes but afford less housing service. To demonstrate how preference heterogeneity in housing service forms a residential sorting, I propose an illustrative model in Appendix to conceptualize the locational decisions by households with heterogeneous preferences for housing. Therefore, I assume that households have heterogeneous preferences for housing and homogeneous preferences for other attributes.

To estimate heterogeneous preferences for housing, I construct a mixed logit model that accommodates random coefficients (McFadden and Train, 2000) and adopts a two-stage estimation strategy to get around the potential bias from omitted locational attributes in estimating the model (Murdock, 2006). The first stage is to estimate a mixed logit model with the systematic utility, V_{ijk}, defined in the equation (1.2). The coefficient, β_1, on housing expense is assumed to be normally distributed, with the mean μ and variance σ^2, i.e., $\beta_1 \sim N(\mu,\ \sigma^2)$. The variance σ^2 is estimated in the first stage, while the means of β_1 are restricted to be zero. The mean of the coefficient β_1, μ, can only be estimated in the second stage, since PUMA fixed effects, η_j, technically absorb all the average influence of locational attributes. In the second stage, I regress the estimated PUMA fixed effects on all locational attributes to estimate the mean coefficients of housing expense and other attributes as follows:

$$\widehat{\eta}_j = \mu \text{Hcost}_j + \lambda' \Phi_j, \tag{1.3}$$

where $\widehat{\eta}_j$ are the PUMA fixed effects estimated in the first stage. Hcost_j is the average housing

expense in PUMA j and Φ_j denote all other locational attributes for which households have homogeneous preferences, including the traffic congestion level that is introduced later. This approach essentially treats the locational fixed effect as a quality of life (QOL) index, which is equal to a weighted sum of urban amenities and other location-specific attributes (Albouy, 2016).

Assuming that the idiosyncratic errors, ε_{ijk}, are independently and identically distributed with Type I extreme values, the probability of household i choosing PUMA j as a residence and j as a workplace becomes:

$$\text{Prob}_i(jk) = \int \frac{\exp(\text{Hcost}_{ij}\beta_1 + Z'_{ijk}\theta)}{\sum_{jk}\exp(\text{Hcost}_{ij}\beta_1 + Z'_{ijk}\theta)} f(\beta_1|\mu,\ \sigma^2)d\beta_1, \tag{1.4}$$

where $Z_{ijk} = [(\text{Income}_{ik} - \text{NonHcost}_{ij}),\ \Gamma_{jk},\ M_{ij},\ \eta_j]$ are the utility determinants with homogeneous preferences. $f(\beta_1|\mu,\ \sigma^2)$ is the probability density function of β_1 that follows the normal distribution. The parameters of the equation (1.4) are estimated by maximizing the following simulated log-likelihood (*SLL*) function (Hole, 2007):

$$SLL = \sum_{i=1}^{N} \text{In}\left\{\frac{1}{R}\sum_{r=1}^{R}\prod_{jk}\left[\frac{\exp(\text{Hcost}_{ij}\beta_{1i}^{[r]} + Z'_{ijk}\theta)}{\sum_{jk}\exp(\text{Hcost}_{ij}\beta_{1i}^{[r]} + Z'_{ijk}\theta)}\right]^{h_{ijk}}\right\}, \tag{1.5}$$

where $\beta_{1i}^{[r]}$ is the r-th draw of the coefficient β_1 for the household i from the normal distribution, $N(\mu,\ \sigma^2)$. R is the number of draws of the random coefficient for each household. h_{ijk} is an indicator variable that equals 1, if the household i selects alternative j and k, and 0 otherwise.

1.6.2 Wage Equation

When making a locational decision, households typically take into account a household permanent income comprised of future income streams, rather than the current income (Friedman, 1957). The expect permanent income the household i can receive if working in PUMA k, Income_{ik}, can be decomposed into a wage income and a non-wage income, i.e., $\text{Income}_{ik} =$

$\sum_w \text{Income}_{iwk} + b_i$. Income_{iwk} is the wage earned by the worker w and b_i is the non-wage income or retirement income of household i that is unrelated to the locational choice. The wage of an earner depends on both the demographics of the working member and local labor market structure. Using the classic two-step instrumental variable approach by Rapaport (1997), I estimate the following wage equation:

$$\begin{aligned}\ln(\text{Income}_{wkm}) \,|\, \langle D_w, k, m\rangle =& D_w'\Pi + \text{CollegeD}_w \times L_m + \text{GradD}_w \times L_m + \text{O}_w \times L_m \\ &+ \ln(\text{pden})_k\beta + \ln(\text{pden})_k \times L_k + L_k + \varepsilon_{wkm},\end{aligned} \tag{1.6}$$

where Income_{wkm} is the wage of worker w working in PUMA k in metropolitan area m.[48] D_w is the vector of the worker's demographics, including age, educational attainment, gender, race, citizenship, and health status. CollegeD_w and GradD_w denote the bachelor's degree and graduate degree for the worker w. They are interacted with L_m, the indicator variables of metropolitan areas m, to capture the potential wage premium on higher education in metropolitan areas. The interaction terms, $\text{O}_w \times L_m$, control for the potential occupation-related wage differential across metro areas. According to the Standard Occupational Classification, this paper categorizes all jobs into 22 regular occupations, excluding military, as shown in Table 1.11 in the Appendix.[49] $\ln(\text{pden})_k$ denotes the natural log of the density of working population in PUMA k and it controls for the labor supply level in the overall market. The interaction terms, $\ln(\text{pden})_k \times L_k$, are incorporated in the wage equation to account for the agglomeration effect on wage premium in each PUMA. L_k are the locational fixed effects controlling for the local labor market structure and other factors that influence wage incomes.

In the second step of the IV approach, using estimated coefficients estimated in the wage equation (1.6), I calculate the alternative wage incomes, $\widehat{\text{Income}}_{iwk}$, in any given workplace a worker may choose. The predicted household permanent income, $\widehat{\text{Income}}_{ik}$, for the household i

[48] All income wages are inflation-adjusted to 2017 U.S. dollars.

[49] http://www.census.gov/people/io/methodology/

working in PUMA k becomes:

$$\widehat{\text{Income}}_{ik} = \sum_{w} \widehat{\text{Income}}_{iwk} + b_i. \tag{1.7}$$

1.6.3 Traffic Congestion Level

As an important locational attribute, the traffic condition in a local economy impacts the desirability of a residential location. To construct an index of a local traffic congestion level, I first estimate the following commute time equation (Duranton and Turner, 2018):

$$\ln(\text{ComTime})_{jkw} = \ln(\text{ComDis})_{jk}\beta_1 + \ln(\text{pden})_{jk}\beta_2 + X_w\theta + G_{jk}\gamma + \lambda_j + \varepsilon_{jkw}, \tag{1.8}$$

where ComTime_{jkw} and ComDis_{jk} represent the commute time and commute distance between j and k for the commuter w.[50] If the residence lies in the same PUMA as the workplace, the commute distance is assumed to be 0, but the commute time is still observed. $\ln(\text{pden})_{jk}$ denotes the natural log of the population density in all areas through which the commute route goes. X_w is the vector of the commuter's demographics, including the log of income, a quadratic term of age, gender, and educational attainment. G_{jk} is the geographic controls that involve the average precipitation, snowfall, and temperature on the route connecting PUMA j and k. λ_j is the locational fixed effect that controls for the influence of the local urban form on commute time.

After estimating the commute time, I construct an overall index of traffic congestion level, Traffic_j, that measures the general accessibility to the surrounding areas of residence j as follows (Caschili and De Montis, 2013):

$$\text{Traffic}_j = \sum_{k} \pi_{jk} f(\text{speed}_{jk}) = \pi_{jj} f(\text{speed}_{jj}) + \sum_{k \neq j} \pi_{jk} f(\text{speed}_{jk}), \tag{1.9}$$

[50] I assume that everyone selects the same commute route between two areas and thus have the same commute distance.

where π_{jk} is the percent of commuters between j and k and π_{jj} is the percent of population who both live and work in PUMA j. $\text{speed}_{jk} = \left(\frac{\text{ComDis}}{\text{ComTime}}\right)_{jk}$ represents the average speed when commuting between j and k. speed_{jj} represents the average commute speed within PUMA j. $f(\text{speed}_{jk})$ is the impedance function that has a typical exponential form, i.e., $f(\text{speed}_{jk}) = 1/\exp(\text{speed})_{jk}$. In order to make the index more readable, the traffic congestion level calculated in the equation (1.9) is then normalized to the range $[0,1]$ by the formula $\text{Traffic}_j^{\text{norm}} = \frac{\text{Traffic}_j - \text{Traffic}_{\text{min}}}{\text{Traffic}_{\text{max}} - \text{Traffic}_{\text{min}}}$, where $\text{Traffic}_{\text{min}}$ and $\text{Traffic}_{\text{max}}$ denote the maximum and minimum traffic congestion index, respectively.

1.6.4 Spatial Housing Price Equation

The last economic component in the economic model system is a local housing market, in which the average housing expense depends on local demographic compositions, local housing supply, and locational attributes. Given the potential spatial dependence in housing expenses and price determinants in a spatial economy, I construct a spatial econometric model that accounts for spatial dependence and has the ability to uncover both direct and spillover effects of locational attributes. A spatial Durbin model (SDM) is well suited to model the influence of residential sorting on housing markets since it includes spatial lags of both the dependent variables and explanatory variables (LeSage and Pace, 2009). I estimate the local housing expense in the following spatial model:

$$\text{In}(\text{Hcost}_{jt}) = X'_{jt}\beta_1 + w'_j[\rho H + X\beta_2] + L_j + \lambda_t + \varepsilon_{jt}, \tag{1.10}$$

where Hcost_{jt} is local average housing cost in PUMA j in year t.[51] $X_{jt}(k \times 1)$ is the vector of all regressors that influence the local housing expense in PUMA j, where k is the number of regressors. Given the data availability, I adopt many variables from multiple sources in X_{jt}, in-

[51] I take the housing expense, instead of housing price, as the dependent variable because households usually have a better sense of the cost of housing based on the actual expenditure, rather than the property value.

cluding the average household income and population density that control for local demographics, housing supply level, urban amenities, such as traffic congestion level, and climatic attributes. β_1 and β_2 are the coefficients representing their direct and spillover effects, respectively. ρ is the spatial autoregressive parameter measuring the existing spatial dependence of the housing expenses. $H(n \times 1)$ is the vector of housing expenses in all n PUMAs and $w_j(n \times 1)$ is jth row in the spatial weight matrix, $W(n \times n)$. The term, $\rho w_j' H$, captures the spatial impacts of housing expenses in all surrounding areas weighted by w_j. The matrix, $X(n \times k)$, includes the regressors for all observations and $w_j' X \beta_2$ measures the spillover effects of price determinants in neighboring areas weighted by w_j. L_j and λ_t are the locational and year fixed effects, respectively. ε_j is the idiosyncratic error. Given the structure formation in the spatial equation (1.10), the reduced form of the model is derived to represent the direct and spillover effects of r-th regressor of the PUMA j on housing expense in PUMA i:

$$\frac{\partial y_i}{\partial x_{jr}} = (I_n - \rho W)^{-1}(I_n \beta_{1r} + W \beta_{2r})_{ij} = S_r(W)_{ij}, \tag{1.11}$$

where $\beta_{1r} \in \beta_1$ and $\beta_{2r} \in \beta_2$ are the coefficients on the r-th independent variable. $I_n(n \times 1)$ is the vector of ones. The diagonal elements of $S_r(W)_{ij}$ are the direct effects. The sum of off-diagonal elements across each row represents the indirect effect of one unit change in r-th independent variable across all spatially correlated observations on the ith housing expense.

The key component in the spatial hedonic housing pricing model is the spatial weighting matrix, W, which is a block diagonal matrix describing spatial influences among local housing markets as follows:

$$W = [w_1, \ldots, w_j, \ldots, w_n] = \begin{bmatrix} 0 & w_{12} & \ldots & w_{1n} \\ w_{21} & 0 & \ldots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{n1} & w_{n2} & \ldots & 0 \end{bmatrix}. \tag{1.12}$$

The diagonal elements of the matrix are set to zero since there exists no spillover effects on itself. The weights, w_{ij}, is first calculated using the most robust double-power distance weighting as:

$$w_{ij} = \begin{cases} \left[1-(d_{ij}/d)^k\right]^k, \ 0 \leq d_{ij} \leq d \\ 0 \qquad\qquad , \ d_{ij} > d \end{cases}, \tag{1.13}$$

where d_{ij} is the distance between two centroids of PUMAs. d denotes the maximum radius of influence (bandwidth) and k is the integer.[52] The calculated weights are then row-standardized, such that $\sum_j w_{ij} = 1$.

This paper assumes that the level of local housing supply is exogenous in the short run and changes only by the policy-related shock. However, the average household income and population density in the housing demand side can be endogenous on account of the reverse causality in the relationship between the housing expense and these attributes. To address the endogeneity, I adopt an instrumental variable approach and estimate a spatial two-stage least square model (S2SLS) under the framework of the spatial Durbin model (SDM). Following the paper by Mussa et al. (2017), I select the associated data in the 2000 U.S. Census as the instrumental variables (IV) in the first stage regression. These lagged variables are relevant to the current values of local population density and average income but unrelated to other unobserved locational attributes that influence the local housing expense.

1.6.5 Estimation Procedure

The procedure of model estimations is introduced below. Given the model specifications of the economic components, I estimate the entire economic model system in the following

[52] I adopt the typical value, 2, and some different values of k are also taken but found to be robust (Dou et al., 2016). To obtain the radius that well describes the range of spatial effects in California, I test a few values and find that the radius of 100 miles can best fit the data in estimating the spillover influence. Some other spatial weight matrices based on the distance are adopted for the robustness check, but the estimation results are very similar.

order, as shown in Figure 1.11. I first estimate the equation of commute time and calculate the traffic congestion levels. Then, the wage equation and housing price equation are estimated. At last, I estimated the household locational choice model. Given the choice set composed of 265 alternative residential locations and multiple nearby workplaces available to each of 737,503 households and 942,411 workers, I need to estimate location choice model over 3 billion observations.[53] To address the large computational burden, some papers adopt the sampling of alternatives that reduces the number of alternatives in the mixed logit model. It has been shown that a sampling strategy can theoretically produce consistent parameter estimates, but it loses some efficiency (Guevara and Ben-Akiva, 2013). By virtue of sufficient computing power in a computer server, the preference parameters in the household location choice model are estimated over the full choice set.[54] The model estimation is executed in PandasBiogeme, a free package for discrete choice modeling (Bierlaire, 2003).

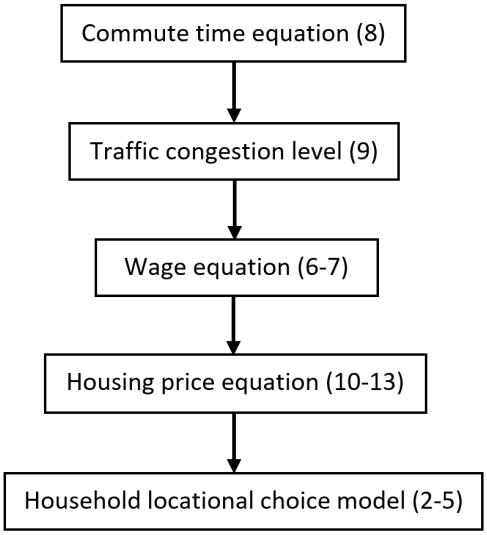


Figure 1.11: Computational Process for the Economic Model System

[53]For each residence chosen by a working household, I select all workplaces within a radius of 70 miles. For retired households, I assume that they can choose from all 265 PUMAs in the complete choice set to settle down. Eventually, there are a total of 3,608,765,900 observations in the estimation.

[54]Specifically, the model is estimated in C5 instance in the Amazon server. It performs in 3.0 GHz Intel Xeon Platinum processors, offering 72 vCPU and 144 GiB of memory.

1.7 Empirical Results

This section presents the empirical results in calibrating the economic model system following the steps in Figure 1.11.

1.7.1 Commute Time and Traffic Congestion

The commute time equation (1.8) is first estimated and Table 1.4 presents the estimation results. It shows that, overall, the local locational attributes, especially the population density, have large impacts on the time spent in a home-to-work journey, while the commuter's personal demographic characteristics barely influence the commute time. Moreover, controlling for fixed effects of a local commuting system largely improves the model fit, showing the importance of the urban transportation system in determining the commute time. Specifically, it is seen that a one percent increase in the commute distance in each route is expected to increase the commute time by 0.8%. The population density has a statistically significant influence on the commute time at 1% level, and the average commute time increases by 0.3% if the local population density rises by 1%, holding other factors equal. It confirms that the city agglomeration leads to an average of more commute time for local commuters. The commuter's income, age, and educational attainment barely have significant influences on the commute time. Female commuters spend an average of 2% more time than males on a trip, *ceteris paribus*. In terms of geographical controls, an increase in annual precipitation and snowfall can result in slower traffic. A warmer winter reduces the commute time, while it takes more time on the route in a hotter summer.

Figure 1.12 shows the average commute speed in each route within the San Francisco metropolitan area. Generally, the traffic is significantly slower in downtown urban areas with a high population density than that in the suburban areas. Given the formula (1.9), I calculate the traffic congestion level for each PUMA based on the average commute time on each route and normalize the index. Figure 1.13 illustrates the geographical profile of the traffic congestion level

Table 1.4: Estimation Results of the Commute Time Equation

Dependent variable: $\ln(\text{ComTime}_{jkw})$				
Variables	[1]	[2]	[3]	[4]
ln(ComDis)	0.8471***	0.7572***	0.7771***	0.7973***
	(0.0123)	(0.0121)	(0.0141)	(0.0118)
ln(pden)	0.2783**	0.2314**	0.3183***	0.3431***
	(0.1212)	(0.1013)	(0.1103)	(0.1120)
Commuter's demographics				
ln(Income)		0.0551	0.0372	0.0409
		(0.1228)	(0.1386)	(0.1048)
Age		-0.0045	-0.0081*	0.0096*
		(0.0228)	(0.0086)	(0.0048)
Age^2		-0.0525*	0.0172	0.0296*
		(0.0228)	(0.0186)	(0.0398)
CollegeD		0.0155	0.0309	0.0316
		(0.1220)	(0.1396)	(0.1389)
Female		0.0125**	0.0180**	0.0196**
		(0.0058)	(0.0089)	(0.0094)
Geographical attributes				
Precipitation			0.0241***	0.0293***
			(0.0006)	(0.0008)
Snow			0.0015**	0.0019***
			(0.0005)	(0.0006)
Summer temp			0.0004**	0.0003***
			(0.0002)	(0.0001)
Winter temp			-0.0021*	-0.0093*
			(0.0012)	(0.0055)
PUMA FE	N	N	N	Y
Observations	1,872,509	1,872,509	1,872,509	1,872,509
Adjusted R^2	0.0821	0.0927	0.1021	0.1820

Notes: This table displays the estimation results of the commute time equation (1.8) over the entire sample. The robust standard errors are presented in parentheses and are clustered at the PUMA level. *** p<0.01, ** p<0.05, * p<0.1.

across PUMAs. It is seen that the level of local traffic congestion is much higher in the centers of metropolitan areas in California.

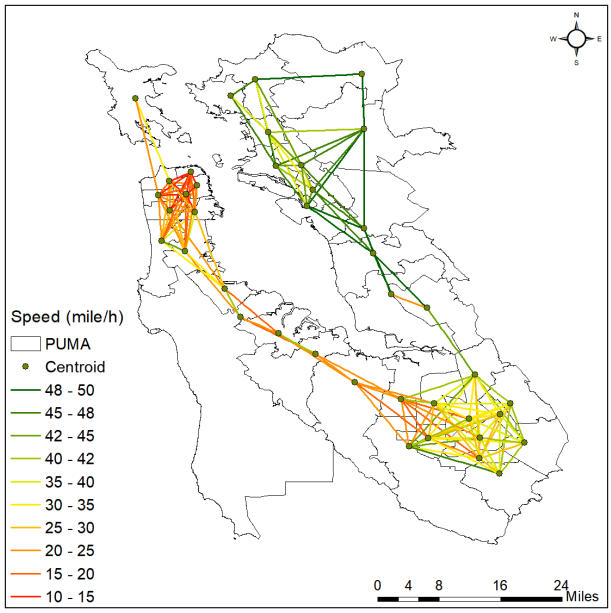


Figure 1.12: Commute Speed in SF Metro Area

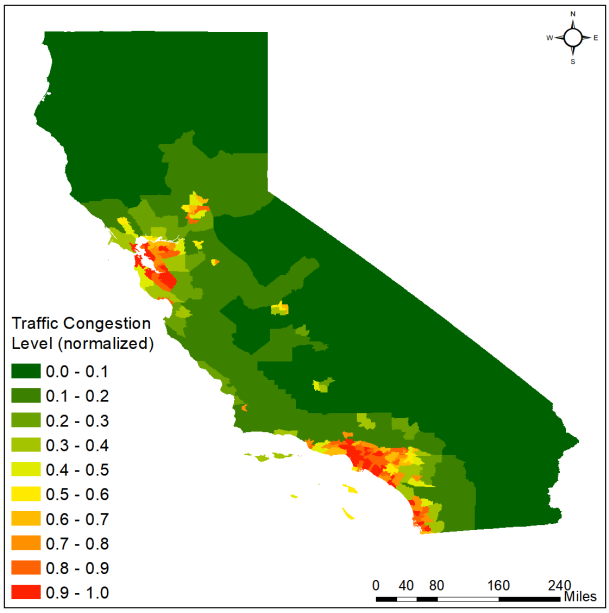


Figure 1.13: Traffic Congestion Level

1.7.2 Wage Equation

Table 1.5 presents the estimation results of the wage equation (1.6). Most coefficients are statistically significant at the 1% level and conform with the conventional wisdom. It shows that, in the labor market in California, a worker with higher education can earn a higher wage than those without it. Citizens and white people stand some advantage in the labor markets, while females and workers with a disability are discriminated against in some sense.[55] Moreover, there exists a nonlinear relationship between the age of a worker and the wage, showing that the personal productivity peaks in middle age.

Apart from the demographics, I explore the wage premiums coming from education and urban agglomeration effects. Holding a bachelor's degree puts a significant premium on a

[55]The defined disability in the census data does not specify the type of physical features, which thus relates to an overall impact.

Table 1.5: Estimation Results of a Worker's Wage Income

Dependent variable: $\ln(I_{wkm}) \mid \langle D_w, k, m \rangle$ in the equation (1.6)					
Variable	Estimate	Std Err	Variable	Estimate	Std Err
Earner's demographics					
High	0.1567***	(0.0031)	CollegeD	0.1039***	(0.0023)
GradD	0.0576***	(0.0024)	Female	-0.0142***	(0.0023)
Age	0.0229**	(0.0102)	Age^2	-0.0003*	(0.0002)
White	0.0130***	(0.0021)	Citizenship	0.0304***	(0.0043)
Disability	-0.0312***	(0.0023)			
Wage premium of education					
CollegeD×LA	0.2033***	(0.0009)	CollegeD×SA	0.0912***	(0.0003)
CollegeD×SD	0.1483***	(0.0004)	CollegeD×SF	0.1945***	(0.0005)
GradD×LA	0.1219***	(0.0007)	GradD×SA	0.0899***	(0.0002)
GradD×SD	0.1115***	(0.0004)	GradD×SF	0.2831***	(0.0011)
Wage premium of urban agglomeration effect					
ln(pden)	-0.0809***	(0.0012)	$\ln(\text{pden}) \times L_k$	Not displayed	
Average coefficients across metropolitan areas					
ln(pden)×LA	0.1433		ln(pden)×SA	0.0812	
ln(pden)×SD	0.1013		ln(pden)×SF	0.1845	
Other control variables					
Occupation × Metro, $O_w \times L_m$					
Locational fixed effects, L_k					
N = 942,411, adjusted R^2 = 0.2519					

Notes: This table displays the main results of the wage equation (1.6) for individual permanent income over the entire sample of the 942,411 workers in the labor market in California. Standard errors are robust standard errors and clustered at the PUMA level, *** p<0.01, ** p<0.05, * p<0.1. The abbreviations of metro areas are as follows: LA (Los Angeles, Long Beach, and Riverside), SA (Sacramento, Arden-Arcade, and Roseville), SD (San Diego, Carlsbad, and San Marcos), and SF (San Francisco, San Jose, and Oakland).

worker's wage and a graduate degree helps further raise the wage in all metropolitan areas. In terms of the local population density, it can be seen that a higher level of population density increases the labor supply in the local market, which decreases the average wage, on the one hand. However, on the other hand, there exist wage premiums in many PUMAs in metropolitan areas due to an urban agglomeration effect. Figure 1.14 presents the PUMA-specific coefficients on the interaction terms, $\ln(\text{pden}) \times L_k$. It shows that, in some highly productive areas, the agglomerations increase productivity, which dominates the impact of a higher level of labor supply and results in wage premiums. Given the estimated coefficients, I predict the alternative wages for each earner if they were working in another workplace.

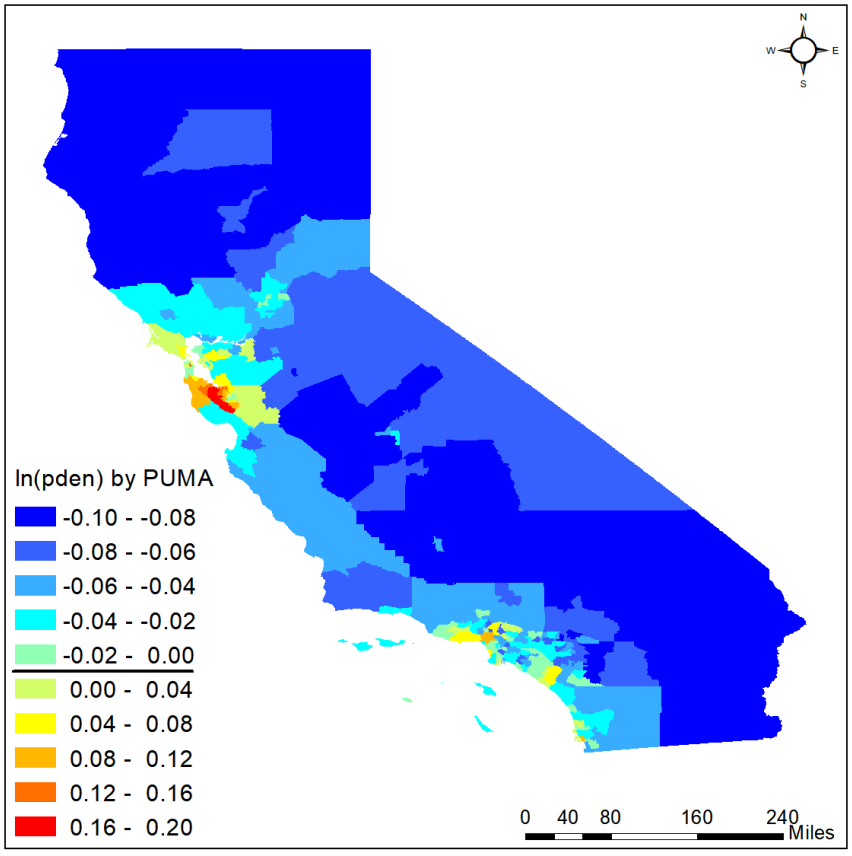


Figure 1.14: Agglomeration Effects on Wage Premium across PUMAs

1.7.3 Spatial Housing Price Equation

This section presents the estimation results of the spatial housing pricing model (1.10) that examines the determinants of the local housing expense. In an attempt to identify the spatial interactions between locational attributes and housing expenses, I first test the possible spatial dependence using an LM test in which a non-spatial hedonic model is assumed to be the null hypothesis (H_0) against a spatial Durbin model (H_1).[56] Table 1.6 presents the test results, showing that the non-spatial models are all rejected at 1% level and there exist strong spatial interactions in the regressors. The LM test statistic in a spatial Durbin model is larger than that in a spatial lag model, which suggests that the spatial Durbin model is preferred and controls for the spatial effects of both housing expenses and locational attributes.[57] Then, I adopt the Durbin-Wu-Hausman test to identify the potential endogeneity in the average household income, population density, and traffic congestion level as a result of the reverse causality (Anselin and Lozano-Gracia, 2008). Table 1.6 shows that the Durbin-Wu-Hausman test statistics are all statistically significant at the 1% level, which rejects the null hypothesis that the three variables are exogenous. The test results justify the selection of the spatial two-stage least square model (S2SLS) under the framework of the spatial Durbin model (SDM).

Table 1.7 reports the estimation results of all housing price models. The column [1] presents the results of the non-spatial housing pricing model. Columns [2-4] show the estimation results of the standard spatial Durbin model, assuming that all regressors, except for spatially lagged housing prices, are exogenous. The standard spatial Durbin model can be estimated using the classic Maximum Likelihood (ML) procedure that yields consistent coefficients. To address the endogeneity, I instrument the endogenous variables using their lagged variables and estimate a spatial housing price model. Columns [5-7] present the estimation results of the spatial two-stage

[56] The non-spatial is the simple OLS model, $\ln(H_{jt}) = X'_{jt}\beta_1 + L_j + \lambda_t + \varepsilon_{jt}$, where the coefficients are defined the same as the spatial housing pricing model (1.10).

[57] The spatial lag model only controls for the spatial lag of the dependent variable, i.e., $\ln(\text{Hcost}_{jt}) = X'_{jt}\beta_1 + \rho w'_j H + L_j + \lambda_t + \varepsilon_{jt}$.

Table 1.6: Tests for Spatial Dependence in the Model Selection

Test	Statistic	p-value
Non-spatial v.s. Spatial		
LM-spatial lag model	54.98	0.000
LM-spatial Durbin model	387.14	0.000
Endogeneity in regressors: Durbin-Wu-Hausman test		
ln(pden)	30.98	0.000
ln(Income)	19.89	0.000
T	18.12	0.000

Notes: LM-spatial lag tests a non-spatial hedonic model (H_0) against a spatial lag model (H_1). LM-spatial Durbin model contrasts a non-spatial hedonic model (H_0) against a spatial Durbin Model (H_1).

least square model (S2SLS). I estimate all the models using the HAC standard errors that allow for the remaining spatial autocorrelation and heteroskedasticity of an unspecified nature (Kelejian and Prucha, 2007).

It is seen in column [1] that, without controlling for the spatial effects, a higher level of the local population density is expected to reduce the local housing expense, *ceteris paribus*. The coefficients are largely biased, and some other coefficients, such as traffic congestion level, T, are also misleading in the non-spatial model, which proves the incompleteness of naïve OLS method. The significance in the autoregressive coefficient, ρ, suggests that there exist spatial interactions among local housing expenses. Given the test results of endogenous regressors, I mainly focus on the estimation results of the spatial two-stage least square model (S2SLS), as reported in columns [5-7]. Compared with the non-spatial model and standard SDM, the S2SLS largely improves the model's fit to data, and its estimates are in line with conventional wisdom. It shows that a higher population density increases the aggregate housing demand in a local housing market. Holding the population density in neighboring areas equal, a one percent increase in the local population density would raise the local housing expense by nearly 0.86%. If the population densities in all nearby areas increase by 1%, the higher levels of aggregate housing demand in the neighboring communities would be spread out and, in the aggregate, fuel the price appreciation by around

Table 1.7: Estimation Results of Non-spatial and Spatial Housing Pricing Models

Dependent variable: $\ln(\text{Hcost}_{jt})$, Hcost_{jt} is the average housing cost in PUMA j in year t

	Non-spatial	Spatial Durbin model (1.10)					
	OLS	Exogenous regressors			Endogenous regressors		
	Direct	Direct	Spillover	Total	Direct	Spillover	Total
Variables	[1]	[2]	[3]	[4]	[5]	[6]	[7]
ρ		0.3917***			0.3298***		
		(0.0242)			(0.0234)		
Local demographics							
ln(pden)	-0.0871**	0.7314***	0.4883***	1.2197***	0.8616***	0.3310***	1.1926***
	(0.0427)	(0.0113)	(0.0223)	(0.0322)	(0.0341)	(0.0291)	(0.0202)
ln(Income)	1.2032**	1.2203***	0.2021***	1.4224***	1.2093***	0.4124***	1.6217***
	(0.0013)	(0.0088)	(0.0016)	(0.0409)	(0.0618)	(0.0110)	(0.0243)
Locational attributes							
ln(hden)	-0.2405***	-0.3253***	-0.2970***	-0.6223***	-0.5760***	-0.1693***	-0.7453***
	(0.0115)	(0.0226)	(0.0143)	(0.0485)	(0.0278)	(0.0240)	(0.0205)
Traffic	0.8902***	-0.7840***	-0.0838***	-0.8678***	-0.4021***	-0.1090***	-0.5111***
	(0.0084)	(0.0670)	(0.0464)	(0.0639)	(0.0674)	(0.1022)	(0.0287)
educ	0.0012**	0.0049*	0.0002	0.0051	0.0055**	0.0036*	0.0091**
	(0.0006)	(0.0027)	(0.0006)	(0.0046)	(0.0027)	(0.0021)	(0.045)
park	0.1102**	0.0940**	0.0014	0.0954*	0.0904**	0.0317*	0.1221**
	(0.0559)	(0.0409)	(0.0031)	(0.0571)	(0.0418)	(0.0177)	(0.0617)
DisCoast	0.0012***	-0.0010**	-0.0002**	-0.0012***	-0.0008**	-0.0001	-0.0009**
	(0.0004)	(0.0005)	(0.0001)	(0.0003)	(0.0004)	(0.0008)	(0.0004)
STemp	0.0012	-0.0009	-0.0004	-0.0013	-0.0008	-0.0004	-0.0012
	(0.0219)	(0.0156)	(0.0237)	(0.0330)	(0.0340)	(0.1083)	(0.0219)
WTemp	0.0014	0.0024	-0.0014	0.0010	0.0080	-0.0014	0.0064
	(0.0216)	(0.0031)	(0.0056)	(0.0168)	(0.0218)	(0.0307)	(0.0330)
precip	0.0032	0.0014	0.0012	0.0026	0.0010	0.0004	0.0014
	(0.0956)	(0.0811)	(0.0246)	(0.0168)	(0.0218)	(0.0389)	(0.0130)
snowfall	-0.0042	-0.0014	-0.0012	-0.0026	-0.0018	-0.0014	-0.0032
	(0.0211)	(0.0313)	(0.0416)	(0.0868)	(0.0203)	(0.0312)	(0.0295)
PUMA FE	Y		Y			Y	
Year FE	Y		Y			Y	
Observations	1,325		1,325			1,325	
Adjusted R^2	0.1827		0.2443			0.3209	

Notes: Columns [1] presents the coefficient estimates of the non-spatial model. Columns [2-3] show the estimation results of the spatial Durbin model with exogenous regressors using ML estimation. Columns [5-7] show the results of the spatial two-stage least square model with lagged variables as instruments for endogenous regressors. The direct and spillover effects are the estimates averaged over all observations arising from changes in each variable. The robust HAC standard errors are presented in parentheses and clustered at the PUMA level. *** p<0.01, ** p<0.05, * p<0.1.

0.33%. The coefficient on ln(Income) implies that home price elasticity of a local average income is 1.21%, and households with a higher income can pay more for the housing service.

Apart from local demographics, the influence of locational attributes is examined. In the S2SLS model, I find that, holding other factors equal, 1% increase in the level of local housing supply can, on average, reduce the local housing expense by about 0.58%. Moreover, locational attributes play an important role in determining local housing expenses. More serious traffic congestion makes the given residence less valuable, while more elementary schools, parks, and being closer to the coast increase the housing price. By contrast, the climatic attributes are found to have little influence on the local average housing expense. It is largely due to the fact that the weather conditions do not vary dramatically across geographical areas in California, and they have a smaller impact on the housing market than other urban amenities.

1.7.4 Household Location Choice Model

Based on the illustration of the simultaneous choice in Figure 1.10, I estimate the household locational choice model in the equations (1.2-1.4). Table 1.8 presents the estimation results of the household location choice model following the two-stage estimation strategy. It includes both the estimates of preference parameters and the marginal willingness to pay (MWTP). I calculate the marginal willingness to pay (MWTP) for each utility determinant by dividing the associated coefficient, β^k, by the coefficient, α, on the Hicksian bundle, $Y-Q$, i.e., $E(\text{MWTP}^k) = -\frac{E(\beta^k)}{\alpha}$. Robust standard errors are reported in the parentheses in the right columns of estimates and MWTPs.

In the first stage, the mixed logit model is estimated with 3,608,765,900 alternatives for 737,503 households. The estimation yields the McFadden's adjusted-R^2 of 0.2769, showing a good fit of the choice behaviors.[58] The model allows the coefficient on housing expense

[58]The formula of McFadden's Adjusted-R^2 is 1-$\frac{LL_m-K}{LL_0}$, where LL_m and LL_0 are log L at convergence and $\beta = 0$, and K is the number of parameters, including the intercept (McFadden, 1978).

Table 1.8: Estimation Results of the Household Location Choice Model

Variables	Estimates (*util*)	Std Err	MWTP ($)	Std Err ($)
The first-stage estimation in the equation (1.2)				
Dependent variable: deterministic utility, V_{ijk}				
Std Dev of β_1 (σ)	0.0190***	(0.0018)		
Income-NonHcost	0.9091***	(0.0228)		
Hcost	-1.2409***	(0.0198)	-$1.36	(0.0217)
Commute distance	-250.55***	(4.4213)	-$0.53/mile	(0.0094)
Commute time	-193.82***	(4.3513)	-$0.41/min	(0.0092)
Move out of PUMA	-1,530.12***	(27.12)	-$1,683.12	(29.83)
Move out of Metro	-478.47***	(28.11)	-$526.31	(30.92)
Moving distance	-8.2364***	(1.0912)	-$9.06/mile	(1.2003)
PUMA Fixed Effects	Y			
McFadden's Adjusted-R^2	0.2769			
Choice observations	3,608,765,900			
Decision makers	737,503			
The second-stage estimation in the equation (1.3)				
Dependent variable: estimated PUMA fixed effects, $\widehat{\eta}_j$				
Traffic	-121.97***	(25.09)	-$134.17	(27.59)
Education	117.78***	(12.07)	$129.56	(13.28)
Park	167.16***	(20.58)	$183.87	(22.64)
Distance to the coast	-25.66**	(2.06)	-$28.23	(2.27)
Summer temperature	-27.25	(20.90)	-$29.97	(22.99)
Winter temperature	12.81	(25.01)	$14.09	(27.51)
Annual precipitation	17.35	22.28)	$19.09	(24.51)
Annual snowfall	-89.20**	(42.94)	-$98.12	(47.023)
R^2	0.1092			
Observations	265			

Notes: This table shows the empirical results of structural estimations on the household location choice model in the equations (1.2-1.4). The marginal willingness to pay (MWTP) is measured by normalizing the coefficients with the Hicksian bundle, Income-NonHcost, i.e., the household income minus the non-housing expense. [1]The coefficient, β_1, on housing expense, Hcost_{ij}, is assumed to be normally distributed, with the mean μ and variance σ^2, i.e., $\beta_1 \sim N(\mu, \sigma^2)$. [2]All MWTPs for commuting costs are measured for one trip, assuming there are 260 workdays and two trips per day. The marginal willingness to pay (MWTP) is measured in 2017 U.S. dollars. Robust standard errors are reported in the parentheses in the right columns of estimates and MWTPs. *** p<0.01, ** p<0.05, * p<0.1.

to be normally distributed, i.e., $\beta_1 \sim N(\mu, \sigma^2)$. The standard deviation of the coefficient is statistically significant, showing that there exists preference heterogeneity in housing service across households. In addition to the statistic of household-specific coefficients, the first-stage model estimation yields the coefficients on the Hicksian bundle, generalized commute cost, and moving cost. It is seen that, on average, a household needs an extra \$1.36 in the disposable income to compensate for an increase in the housing expense by \$1, suggesting that households are typically more sensitive to the housing cost than the household income. When making a trade-off between commute cost and wage income, it is estimated that households are willing to pay \$0.53 to commute one mile less for a work trip.[59] The implicit value of commute time is estimated to be nearly \$0.41/min, implying that households are willing to pay \$0.41 to save one minute in the commute. In terms of the generalized moving cost, I find that the estimated psychic costs of moving out of a PUMA and a metropolitan area in which a household lived previously are \$526 and \$1,683, respectively. For the economic moving cost, a household would ask for an average of \$9 to compensate for moving one mile further in a relocation.

As a critical component for the household locational choice, the estimated PUMA fixed effects, $\widehat{\eta}_j$, measure the overall impact of all local locational attributes and thus offer much information on the level of desirability for an alternative location (Albouy et al., 2016). The PUMA fixed effects pick up the overall impact of locational attributes, including weather, urban facilities, traffic conditions, climatic attributes, education (Ries and Somerville, 2010), and distance to the coast, etc. Those attributes account for motives of non-economic migration (Albouy, 2008). Figure 1.15 shows the estimated economic values of urban amenities across geographical areas in California.[60] The estimates of PUMA fixed effects take the PUMA of Death Valley as the reference area (dashed area).[61] It is seen that the geographical areas that provide better urban amenities are mainly located in metropolitan areas. The PUMA of the central coastal

[59] Assume there are 260 workdays and two work trips per day for each worker. All MWTPs for an alternative commute mode are calculated for one trip.

[60] The estimates of all MSAs are available from the author upon request.

[61] The area is basically a valley in the desert, located in Eastern California.

area in the west San Diego County is found to have the highest economic values of locational attributes and households are willing to pay up to $8,251 for more desirable urban amenities, compared to the reference area, controlling for the household income and housing expense.[62]

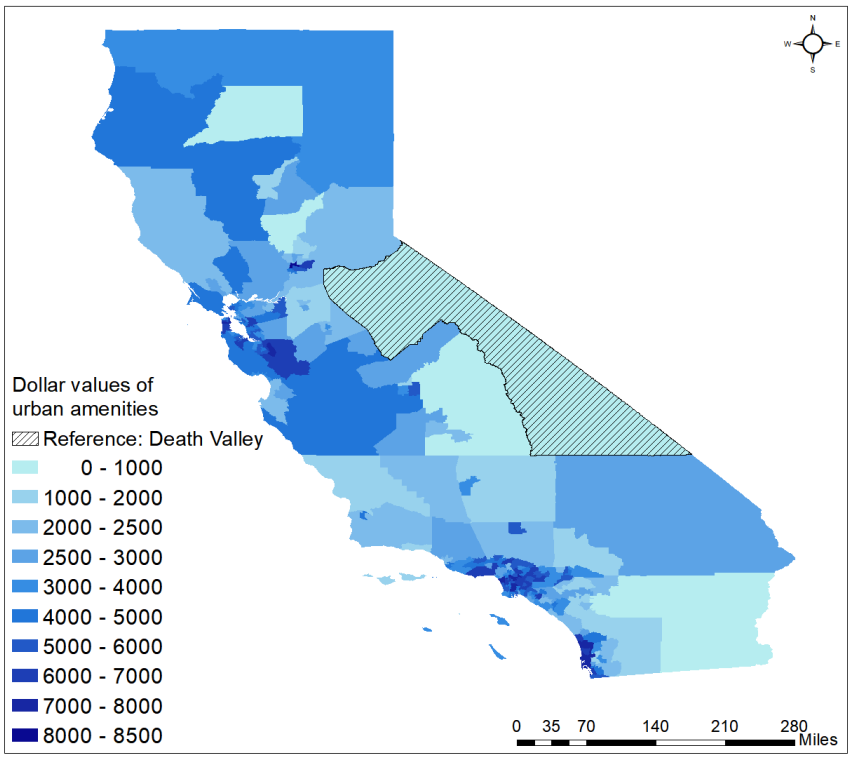


Figure 1.15: Economic Values for the Quality of Life across PUMAs

In the second stage, the estimated PUMA fixed effects are regressed on climatic and locational variables. The bottom panel in Table 1.8 reports both coefficients and MWTPs on these local attributes. It shows that households view a lower level of local traffic congestion as a preferred local amenity. On average, they are willing to pay nearly $134 for the decline of local traffic congestion level by 1%.[63] The estimation result supports the argument that accessibility is an important attribute that determines how desirable a residence could be (Caschili and De Montis, 2013). Moreover, households prefer more elementary schools and public parks in the local area and living in residence closer to the coast. Apart from the urban facilities, the second-stage

[62] The high dollar value is largely due to its friendly climate, coastal view, and better urban facilities.

[63] Since the traffic congestion level is normalized to the range $[0, 1]$, the value equals the improvement of traffic condition by 1%.

estimation yields the coefficients and MWTPs for climatic attributes. Generally, households favor a cooler summer and warmer winter, even if they put little weight on the average temperatures. Empirical results also show that households view a higher level of precipitation as a valuable natural amenity, while a higher level of snowfall is taken as a disamenity. These estimated preference parameters for locational attributes conform to the features of life in California.

Given the estimation results of the household location choice model, I calculate the location-specific deterministic utilities for each of the 737,503 households and evaluate the predictive performance of the model by its in-sample prediction ability. The ability is calculated by the percent correctly predicted (PCP), i.e., the percentage of observations in which the actual choice outcomes correspond to the alternative with the highest probability. Given 265 alternative locations, the PCP is adjusted to calculate the percent of observations in which the chosen location belongs to the ten most probable alternatives. The adjusted PCP is 84%, suggesting that the residential location choice model can accurately predict the household locational choices.

1.7.5 Preference Heterogeneity in Housing and Locational Choices

Heterogeneous Preferences for Housing

When making a locational decision on the residence, household housing preferences can play a systematic role, as illustrated in Appendix. According to the estimation results in Table 1.8, the preference parameter on housing service is found to vary across households and follow the normal distribution, $\beta_1 \sim N(-1.2409,\ 0.0190^2)$. Conditional on the household location choice, h_{ijk}, and observable household demographics and locational attributes, Z_{ijk}, the conditional distribution of the random coefficient on housing preference, β_{1i}, can be derived using the Bayes rule for each household i (Revelt and Train, 2000). Then, for the practical purpose of the estimation, the household-specific housing preferences are simulated, as introduced in detail in Appendix.

Table 1.9: Heterogeneous Housing Preferences by Demographic Groups

Distribution of housing preference parameter: $\beta_1 \sim N(-1.2409,\ 0.0190^2)$					
	Group	Estimate		Group	Estimate
Age of a householder	15-25	-1.6409	Household size	1	-1.4339
	26-35	-1.5293		2	-1.2201
	36-45	-1.4193		3	-1.1903
	46-55	-1.2402		4	-1.2312
	56-65	-1.1391		5	-1.2013
	66-75	-0.9124		6	-1.1927
	Over 76	-0.7890		7	-1.1909
	Group	Estimate		Group	Estimate
Household income	\$2,000-\$30,000	-1.0934	Ethnicity	White	-1.2409
	\$30,000-\$70,000	-1.3087		Non-white	-1.2409
	\$70,000-\$130,000	-1.3493	Marital status	Married	-1.0890
	Over \$130,000	-1.0492		Unmarried	-1.4653

Notes: This table presents the estimated household housing preferences across demographic groups, categorized by the age of a householder, household size, household income, ethnicity, and marital status of household members. Married implies that there is a married couple in the household, and Unmarried means there are only unmarried partners or one single person in the household.

Household housing preferences can be influenced by household demographics. Table 1.9 reports the means of the conditional housing preference parameters that are averaged across all households in each subgroup divided by age, household income, household size, ethnicity, and marital status. A lower value of the housing preference parameter suggests that this demographic group is more sensitive to the housing expense and spends a larger share of household income on housing service.

It can be seen that there exists a positive relationship between the coefficient on housing expense and the age of a householder across age groups. It indicates that younger households are sensitive to the housing expense and are willing to pay a higher portion of their income on the housing service. It is due to the fact that younger residents are mobile and thus have a higher propensity to adjust housing choices. As the head of a household becomes older, a household is less likely to move and allocates a lower budget share to housing, partially because they need to spend more money on retirement life and medical care. In another aspect, the household income

has a significant influence on the housing preference, and there exists a complex relationship between income level and the sensitivity to housing expense. Low-income and high-income households are found to be less sensitive to housing expenses than middle-income households. It also implies that rich and lower-income households pay a smaller portion of household income on housing service.[64] Apart from the age of a householder, the influence of household size on the housing preference is examined. It is found that household size has no significantly systematic influence on the responsiveness to housing expenses, except for the one-person households with the lowest coefficient. Since a majority of one-person households are younger, they are more likely to move in response to the increase in housing expenses. In addition, the housing preference has a consistent distinction by marital status in household members. Households with a married couple are less responsive to a change in housing expense than unmarried partners, which can be largely explained by a more stable living arrangement due to the marriage. As an important demographic factor, ethnicity plays a critical role and has a large predictive power in housing preference, since some deep preferences for housing can arise from ethnicity-related influences, such as culture similar to their birthplace and habit persistence in housing choice. I find that, compared to non-white households, white households spend a larger share of household income on housing service, which is partially due to the fact that white households have a significantly higher WTP for a house with a larger residential space. The consistent large differences in housing preference across demographic groups would play a critical role in household locational choices, forming a taste-based residential sorting.

Heterogeneous Locational Preferences

Given the estimated preference parameters and household demographics, I estimate the probability distribution of locational choices for households using the choice probability equations (1.4-1.5). The heterogeneous preferences for locations across demographically identifiable groups

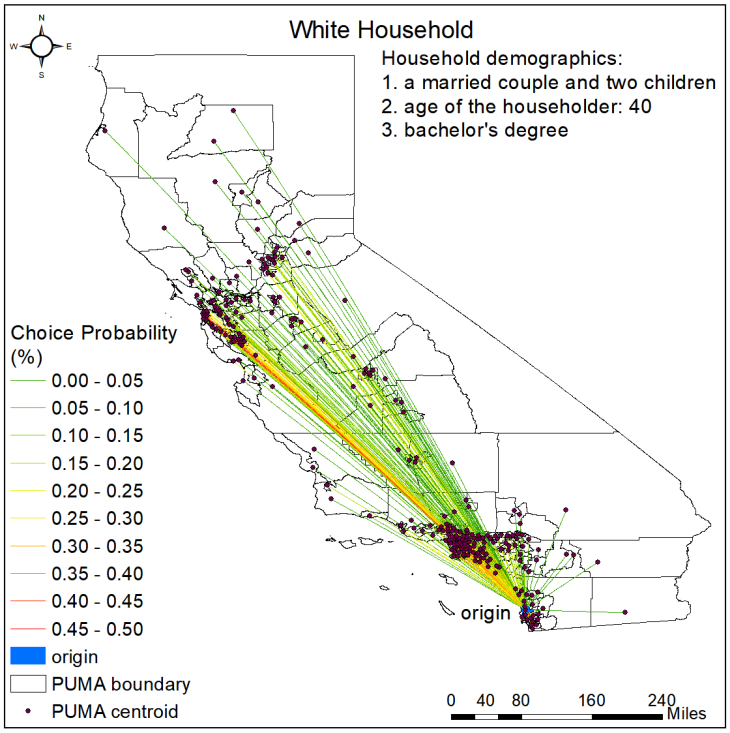
[64] When a household has a low income, a large share of income is spent on other non-housing essentials. For rich families who have sufficient housing service, other non-housing goods and services, e.g., luxury goods, are preferred.

are investigated in the residential sorting to describe the migration pattern.

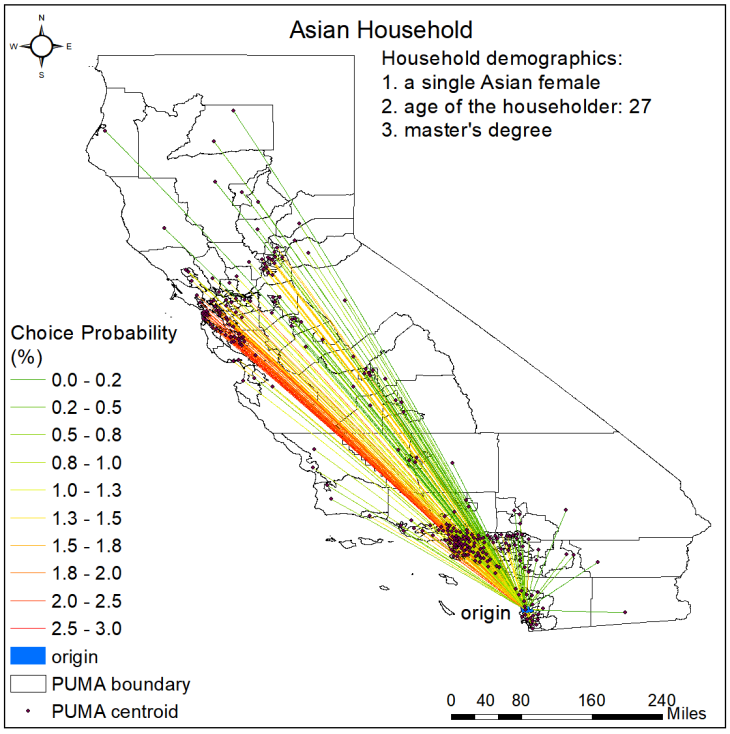
To visualize preference heterogeneity in locations across demographical groups, two households with distinct household demographics are taken as a comparison. Figure 1.16 presents the geographical distributions of probabilities on each PUMA for the two households. The first is a typical white household with a married couple and two children. The male householder, at the age of 40, has a bachelor's degree.[65] The other household is comprised of a single Asian female at the age of 27 with a graduate degree. To make the distributions of choice probabilities more comparable, it is assumed that two households have decided to move from the same origin, University City in the San Diego metropolitan area. For graphically simplicity, I aggregate the choice probabilities of alternative workplaces nearby on all PUMAs to predict the likelihood of moving to each of the 265 PUMAs.

It shows that the two households have very different probability distributions for locational choices. The color of the lines connecting sending and receiving PUMAs represents the choice probability. Both households have a strong preference for staying in the same metropolitan area. In panel (a), it can be seen that the white household has a more spread distribution of choice probabilities across the areas. Some popular urban cities in San Francisco and Los Angeles are given priorities with likelihoods higher than 0.4% in the locational choice. On the contrary, there exists a different migration pattern for the Asian household. As denoted by the red lines, she has a much stronger preference for large urban areas and possibly settles down in these areas with the probabilities greater than 2% in each PUMA. The differing migration choices can be largely interpreted by the wage premium on metropolitan areas and lower costs for housing and non-housing service. This well-educated single young Asian person has a higher income in urban areas but a lower demand for a large residential space. Thus, after consuming sufficient housing, she has a higher portion of the disposable income spent on the non-housing service in an urban area.

[65] To aggregate the information of occupations, the choice probabilities for the two households are calculated with different occupations first and then weighted-averaged by its occupational distribution.



(a) Household 1: a White married couple with two children



(b) Household 2: a single Asian female

Figure 1.16: Maps of Probability Distributions on Locational Choices for Two Households

The comparison between two typical households conforms to the existing migration pattern and provides the utility-consistent evidence for the population migration formed by numerous household relocations. Gradually, the expensive metropolitan areas are inhabited by more educated households who demand less housing service per household member. This long-lasting trend can boost the local housing expenses and make the housing affordability issue more serious in expensive urban areas.

1.8 Economic Consequences of Expanding Housing Supply

Given the calibrated economic model system and household preference parameters, this section explores the economic consequences of expanding the local housing supply in some urban areas. I first examine how effective the proposed housing policy is in solving the housing affordability issue. Then, I examine the changes in local locational attributes, e.g., traffic congestion level, and investigate how the policy would affect the welfare of local residents and new immigrants in these places.

1.8.1 Simulation Procedure

Figure 1.17 presents the simulating process of a one-time shock in the local housing supply. The housing supply expansion first goes into effect in the housing market and lowers the housing expenses in the short run. In response to the new price parameters, households who are free to relocate in an open economic system would possibly choose a new residence to remaximize the household utility. Numerous locational decisions contribute to the residential sorting and thus change the population density in a local economy. As a consequence, the labor market and local commuting system would be affected, and the urban agglomeration effects are reflected in wage premium and traffic conditions. Then, given new locational attributes and household incomes, the local housing expense would be updated and later taken into account

when making a household locational choice once again. The economic model system continues the iteration process and iteratively solves for itself, until it comes to a new equilibrium where households eventually settle down and stop moving.[66]

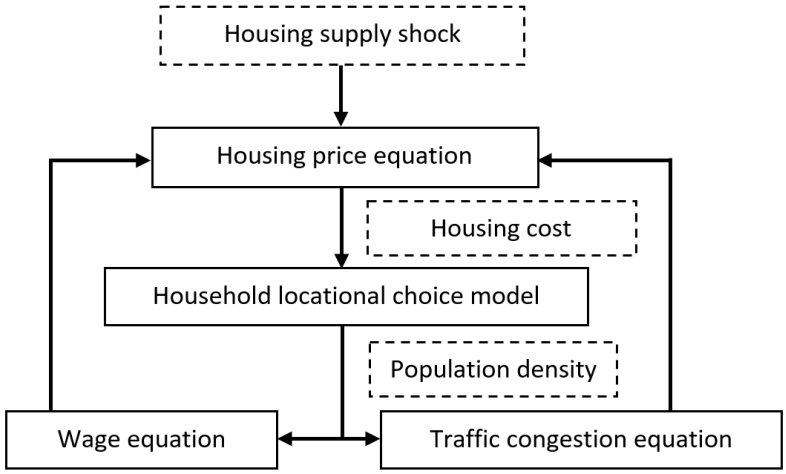


Figure 1.17: Simulation Procedure of a Housing Supply Shock

1.8.2 Simulation Results

To better illustrate the economic consequences of an urban housing supply expansion, I select a representative high-productivity urban area with a serious housing affordability issue facing local residents. Figure 1.18 shows the map and location of the PUMA, comprised of the urban cities Mountain View, Palo Alto, and Los Altos in Northwest Santa Clara County in San Fransisco metropolitan area. This local urban area is an employment center, and there exist many high-tech companies that offer high-paying jobs. The median annual household income is around $124,000. However, many local households, especially those unfortunately put in an economic disadvantage, afford much less housing service than high-income families, and the median monthly housing expense is as high as $4,700, nearly 46% of their household income. Moreover, the local traffic congestion is a serious problem that bothers many local commuters.[67]

[66]It is assumed to reach a new equilibrium if no more than 0.1% of the households living in California are still moving.

[67]The average commute speed in the local area is 18.3 miles per hour.

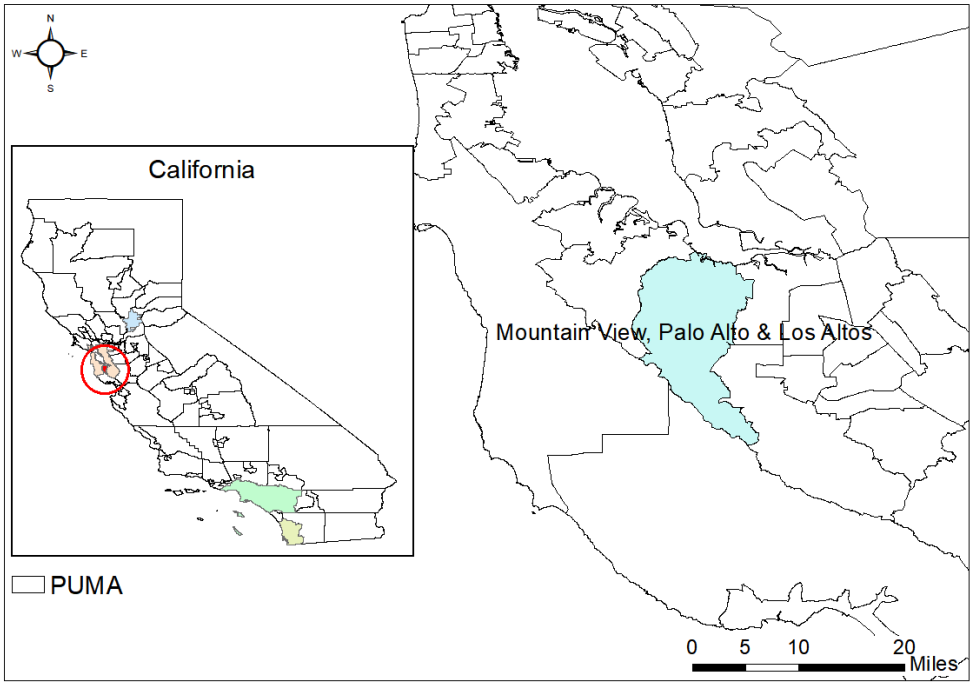


Figure 1.18: Map of the Local Urban Area for Simulation

Now consider the one-time shock in the housing supply side due to the housing policy. The local city government expands the housing supply by 5% in the local housing market.[68] According to the transmission mechanism in Figure 1.17, the economic system would respond and evolve until it reaches a numerical equilibrium. Figure 1.19 shows the dynamic changes in housing expense, population density, household income, and traffic congestion level in the PUMA in the San Fransisco metropolitan area as an example. It is seen from the numerical solution that it takes about 20 iteration steps for the economic model system to converge to a new equilibrium over the relatively short time frame.[69]

In the first iteration, the one-time shock due to the sudden increase in the number of dwelling units drops the expense of housing services. The lower housing expense draws new

[68]The Senate Bill 50 aims to increase the housing supply by nearly 30%. Since it is a bit impractical and dramatically change the housing market structure, I reduce the magnitude of the policy-related shock and adopt 5% to attain more accurate simulation results.

[69]Theoretically, the length of the entire time period depends on how long it takes for each household to relocate in each iteration step. In practice, due to the almost complete information, households can largely forecast the eventual market equilibrium and make the final locational decisions, which makes the convergence of the new equilibrium much quicker.

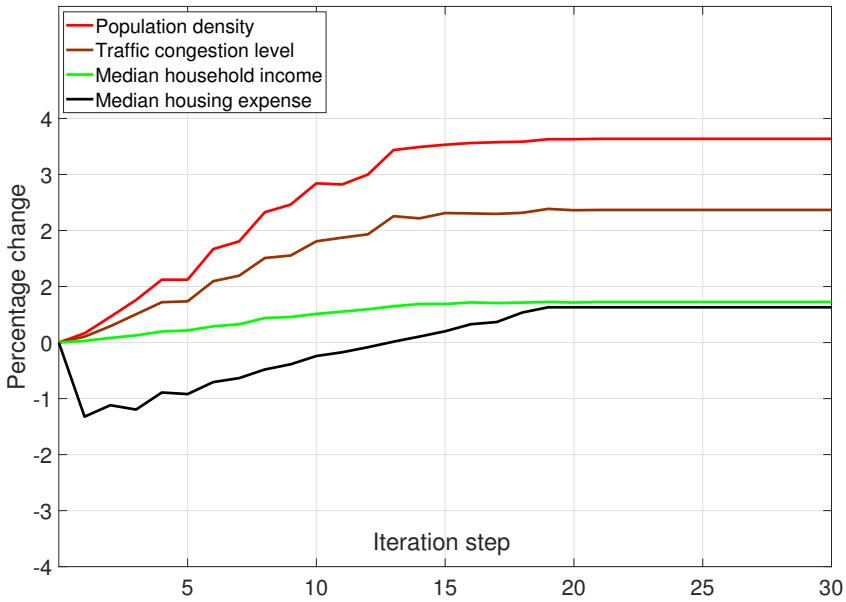
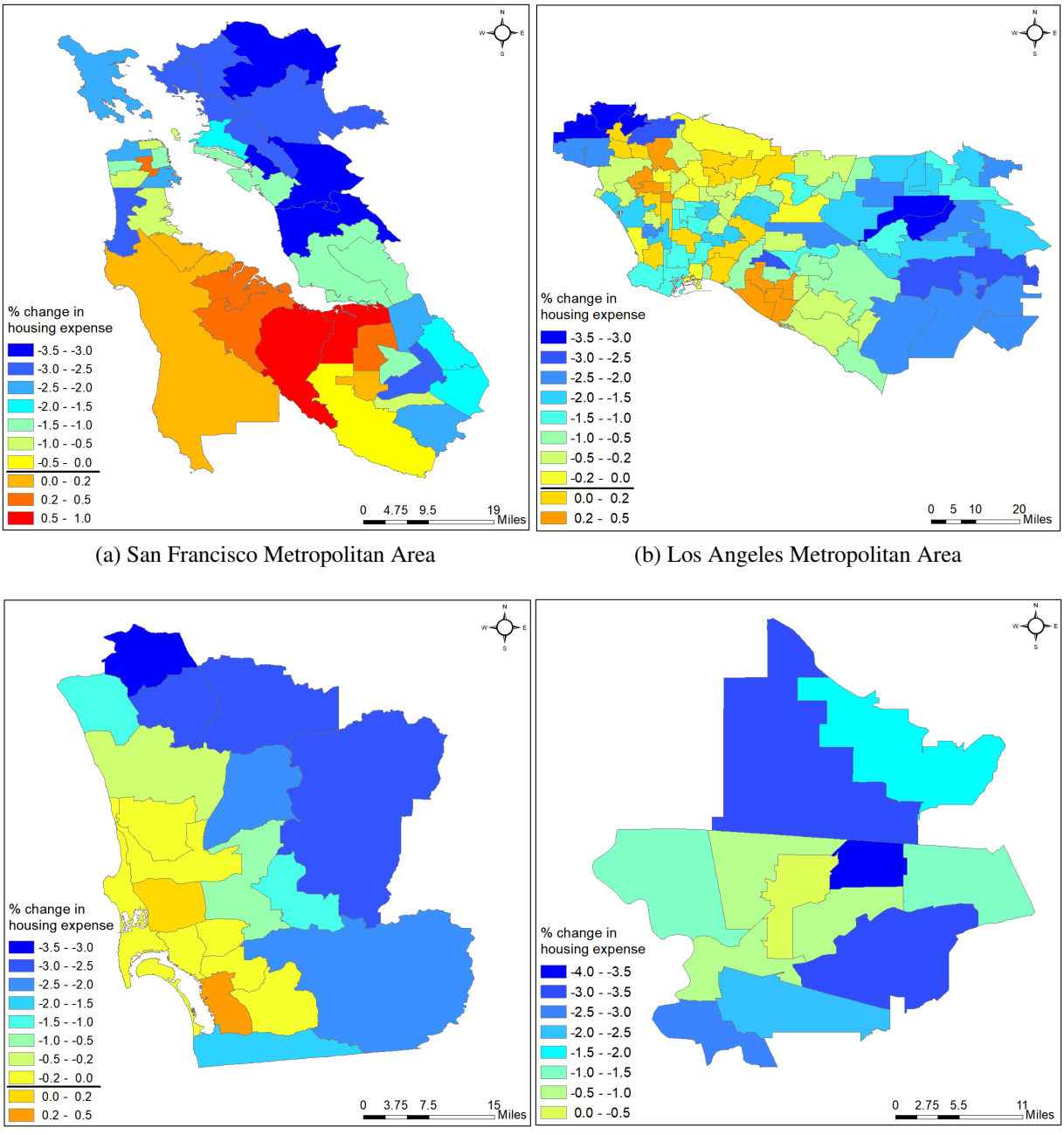


Figure 1.19: Dynamics of the Local Economy

households from outside the high-productivity urban area, which increases the local population density. Over time, the median housing expense gradually goes back up and eventually becomes nearly 1.2% higher than it was were before the one-time shock. This massive in-migration to the productive city, coupled with urban agglomeration effects, also increases income opportunities, and the median household income increases by around 1.4%. Therefore, in an open economic system, the residential sorting can negate much of the one-time housing supply expansion and largely undo what the housing policy aims to achieve. In addition to its ineffectiveness, the housing policy results in some urban disamenities. It is seen that, over the same time period, the local population density has been dramatically increased. The higher local population density causes local traffic congestion to become even worse than before the housing policy.

Apart from the productive area in San Fransisco, using the estimation results, I simulate the economic system model and explore the economic consequences of a 5% increase in the housing supply level in other local housing markets. Figure 1.20 presents the eventual percentage changes in the median housing expense across PUMAs in a metropolitan area if the 5% housing supply expansion is deployed in the given PUMA. It is seen from the simulation results that, if



(a) San Francisco Metropolitan Area

(b) Los Angeles Metropolitan Area

(c) San Diego Metropolitan Area

(d) Sacramento Metropolitan Area

Figure 1.20: Percentage Changes in Median Housing Costs

the housing supply is expanded by 5% in the local housing market, a drop in the housing expense is realized in a majority of the urban areas, except for a few employment centers in San Francisco, Los Angeles, and, San Diego metropolitan areas. These urban areas are productive enough to drain more high-income workers, which boosts the local housing price and completely negates the housing policy.

The real world, specifically the situation in California, is very complex. In an open system, cities with more successful agglomeration effects attract people from less successful cites. If SB 50 were successful in substantially reducing housing prices by building sizable numbers of additional dwelling units in the areas targeted, then two things would happen in these urban areas. The initial impact would be similar to the phenomena that hit Las Vegas, Phoenix, and San Bernardino in the late 2000s, large scale defaults by recent home buyers as the value of their property fell below what they owed mortgage lenders. That would be followed by large inflows of new households, mostly from outside of California, seeking to take advantage of high wages now coupled with lower housing costs. This inflow of new workers would drive down wages exacerbating default rates among recent home buyers. This inflow would, over time, largely undo SB 50's stated purpose of reducing the cost of housing services. Further, as density increased in the longer run, firms in coastal cities in California's most productive cities would invest in capital and create new jobs. It would tend to increase effective wage rates in the effect of an agglomeration economy, which could drive housing costs higher than before the expansion (Ellison et al., 2010).

1.8.3 Aggregate and Distributional Welfare Effects

The framework under which economists have used for decades to evaluate a housing policy is the measurement of welfare in the entire economy. The key feature of an open system of cities is that people maximize their welfare by choosing where to live, taking into account their budget constraints, and the transaction costs associated with moving. Given the calibrated

economic model system, I examine the aggregate and distributional welfare effects of a local housing expansion in California, taking the same PUMA as the example in Figure 1.18 for two reasons. First, Senate Bill 50 is more likely to be passed and implemented in urban areas with a local housing affordability issue, such as the chosen PUMA. Secondly, given the high employment and residential density, the housing policy has a larger impact on the local and entire economy.

According to the estimation results of household preference parameters in the spatial economy, I calculate the household-specific utility in the specification (1.2), conditional on their current locational choices, and use the measurement as a proxy for household welfare. Then, taking into account the positive housing supply shock by 5%, I simulate the economic model system and measure the new household utility in the numerical equilibrium with possibly new locational decisions across households. Given the two utility levels before and after the policy shock, the percentage changes of household welfares are estimated for each household. Given the locational decisions and relocations, I divide the entire population into several categories in relation to the chosen PUMA, the local residents that always stay in the local area, the in- and out-migrants of the PUMA, the households living in San Francisco metropolitan area, and all households who stay in California.

Table 1.10 summarizes the changes in the long-run welfares by both geography and demographic groups due to the housing policy. The changes in household welfare are aggregated and averaged by geography and demographic attributes. It is seen that, from the perspective of geography, local residents living in productive cities are worse off, which can be largely explained by the rising traffic congestions. SB 50 is unrealistically optimistic about the role of public transportation. Public transit use in the San Francisco metropolitan area is quite low with less than 5% of the population, according to statistics by 2017 National Household Travel Survey.[70] The provision of more affordable housing services and resulting higher local population density are guaranteed to produce substantial increases in traffic congestion and are likely to reduce other

[70] https://nhts.ornl.gov/

Table 1.10: Welfare Effects of the Local Housing Expansion with Full Mobility

	Local residents	In-migrants	Out-migrants	San Francisco	California
Education					
No high school	-6.71%	1.12%	-6.82%	-0.95%	-0.013%
High school	-5.70%	2.44%	-6.98%	-0.34%	0.003%
College degree	-2.12%	6.84%	-6.48%	0.41%	0.007%
Graduate degree	-1.26%	11.72%	-6.76%	0.71%	0.011%
Labor participation					
Dual-worker	-1.32%	8.11%	-6.22%	-0.11%	0.015%
One-worker	-5.47%	4.25%	-5.62%	-0.27%	0.005%
Retired household	-2.62%	3.91%	-7.12%	-0.32%	-0.005%
Age					
< 30	-2.62%	9.72%	-5.67%	-0.15%	0.004%
30-65	-4.61%	8.53%	-5.36%	-0.23%	0.005%
> 65	-5.72%	3.44%	-6.52%	-0.31%	0.006%
Household income[1]					
$< 25\%$	-6.81%	8.62%	-6.42%	-0.11%	0.003%
25-50%	-5.42%	7.94%	-5.94%	-0.25%	0.004%
50-75%	-4.71%	6.87%	-6.45%	-0.17%	0.006%
$> 75\%$	-3.72%	4.61%	-7.13%	-0.23%	0.004%
All	-4.40%	6.71%	-6.52%	-0.22%	0.005%

This table presents the welfare effects of the local housing expansion that happens in the chosen PUMA. Household-specific changes in welfares are measured with the household utility specification (1.2) with full mobility. The changes in household welfare measurements are aggregated and averaged by geography and demographic attributes. [1]Household incomes of in-migrants are measured at destinations, and these of out-migrants are estimated at origins using the wage equation (1.6).

quality of life indicators.[71] The beneficiaries of the local housing expansion are those who move into the chosen PUMA. However, the increases in the welfare of in-migrants come at the cost of the welfare of those moving out of the productive cities. Generally, it is estimated that the housing policy harms the public in San Francisco metropolitan area, while the Californians, as a whole, gain welfares at the state level, even if the percentage change is relatively small. Figure 1.21 shows the average percentage changes in household welfares across geographical areas in San Fransisco metropolitan area due to the housing expansion. It shows that the housing supply expansion sort of backfires and does large harm to expensive urban areas. However, the suburban areas and less developed cities nearby benefit from the housing expansion in the chosen PUMA. It is partially due to the fact that many households are attracted to move to expensive and productive cities, which lowers housing expenses and raises the average wage incomes given a lower level of local labor supply.

Apart from geographical variations, I find large variations across demographic groups. In terms of educational attainment, it is seen that well-educated households benefit more from moving into the productive urban area, due mainly to the wage premium on education. However, the less-educated households are put at a disadvantage in the welfare redistribution. As to labor participation, the local dual-worker households lose less welfare than one-worker households, since they benefit relatively more from the wage premium due to the higher local population density. Among the in-migrants, we observe more welfare gains in dual-worker households, largely because they have the opportunity to achieve a good wage income by working in the urban cities. Moreover, the housing policy has a smaller impact on younger local households than older ones, for that younger households are more likely to move and adjust the housing and locational choices to avoid substantial welfare loss. Lastly, I divide the surveyed population by household income. It is found that the housing expansion causes more harm to the local economically

[71] To see why this will happen, consider a simple example which SB 50 wants to see replicated on a large-scale tearing down small single-family houses with a medium-size lot and building small apartment buildings. This increase in residential density from the infill project leads to potentially more fight over increasingly scarce parking spaces.

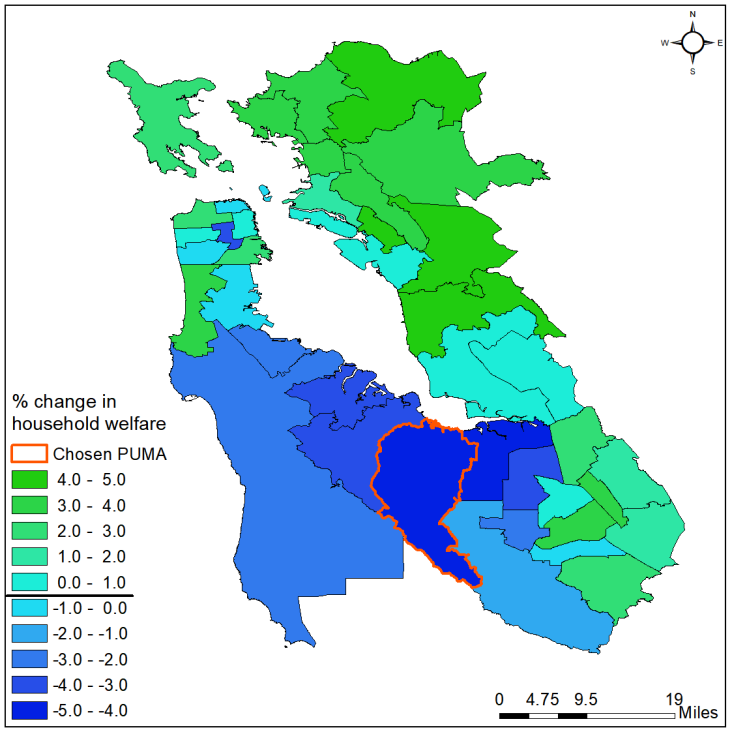


Figure 1.21: Percentage Changes in Household Welfare in San Fransisco Metropolitan Area

vulnerable households than high-income households, while the lower-income in-migrants gain more opportunities and welfares in the labor market.

In sum, the population decomposition by demographic attributes and locational choices shows that the housing expansion policy has different impacts across groups and geographical areas. The changes in wage premiums and quality of public goods, such as traffic congestion, are the most important channel from which the overall welfare redistribution has been formed. For existing local residents, the massive in-migration raises the local population density, which increases the aggregate housing demand and worsens the traffic jam. The benefits natives gain from high productivity due to the urban agglomeration effects cannot offset the slightly higher housing expense and the lower quality of urban facilities. On top of the welfare loss in the local population, the well-educated, younger, and dual-worker households who move into the urban area gain large welfares. As a whole, the entire metropolitan area experiences a welfare loss. However, the local housing expansion can benefit the average household in California through

large-scale population relocation. These results suggest that more desirable urban policies should be implemented to accommodate an even larger population while sufficiently compensating the local residents.

1.9 Conclusion

This paper examines the economic consequences of a proposed housing legislative act that aims to fix the housing affordability crisis by expanding the housing supply in productive urban cities. It also explores how residential sorting plays a role in forming a new market equilibrium. Using the newly released 2013-2017 ACS data, I construct an economic model system that characterizes household residential location choices and their spatially simultaneous interactions with local labor and housing markets and urban amenities across geographical areas in California. The empirical results show that, in an open-city economy, an expansion in urban housing supply leads to large-scale in-migration of well-educated, younger, and dual-worker households to expensive and productive urban cities over time. Due to positive agglomeration externalities in wage premiums, the positive residential sorting would further raise household incomes and largely undo what the housing expansion aims to achieve. In addition to the ineffectiveness of the housing policy, a denser population further exacerbates local traffic congestion and causes a large welfare loss on existing local residents. It is also found that households prefer living in a place where they can attain a higher wage income, pay a lower housing expense and commute cost, and have access to high-quality urban amenities, like less traffic congestion.

The findings of this study advance the understanding of the spatially interdependent relationship between population migration and the local economy and provide substantive implications for regional urban planning. The main takeaway from this paper is that we should carefully examine the behavioral and economic responses, especially unintended consequences when making an urban housing policy. The considerable welfare loss in local urban natives

arises mainly from the worse traffic congestion, suggesting that an urban policy that improves the commuting system needs to be made along with a housing supply expansion. Given the limited capacity in the private commuting mode in dense urban cities, the city government should improve the efficiency of the local public transit system. The proximity to a public transit system has been shown to improve the quality of life and increase the desirability of a residence (Chen and Haynes, 2015). In many modern cities, transit-oriented housing stands as one of the most promising mechanisms for promoting multiple urban policy objectives, including affordable housing construction, sprawl containment, and reduced car dependence (Cervero, 2016). Increasing the capacity of the existing public transit systems and spreading its network, such as Bay area rapid transit (BART) in San Fransisco metropolitan area, can largely mitigate the traffic congestion in the dense urban areas.[72] Therefore, the local government actively participate in some major public investments for infrastructure, ranging from major enhancements to the road system to the heavy rail system, to maintain or improve the quality of urban life.

Apart from the improvement of public urban facilities, public policies related to welfare redistribution should be cautiously made. The purpose of good urban planning is to take account of its cumulative impacts on the entire population in the locality. Under the proposed housing legislation, real estate developers have no liability for the imposition of welfare loss on a substantial fraction of local residents. In the infill areas, it is not hard to forecast that commute time by private cars would dramatically increase due to the change imposed by the government of California on local zoning and that the availability of street parking for both residents and guests would largely disappear.[73] As a result, from the perspective of social welfare, a housing policy of supply expansion should include mandatory impact fees for providing the necessary infrastructure or compensation to fix the problem. The local government or real estate developers

[72]The Bay Area Rapid Transit (BART) is a rapid transit public transportation system serving the San Francisco Bay Area in California. The heavy rail elevated and subway system connects San Francisco and Oakland with urban and suburban areas in Alameda, Contra Costa, and San Mateo counties. See: https://www.bart.gov

[73]Either these folks are social do-gooders who care about poor people but do not know how the housing and labor markets work, or they are convinced by real estate developers who are trying to get properties up zoned since tearing down a single-family house and building apartments can dramatically increase the value of urban lands.

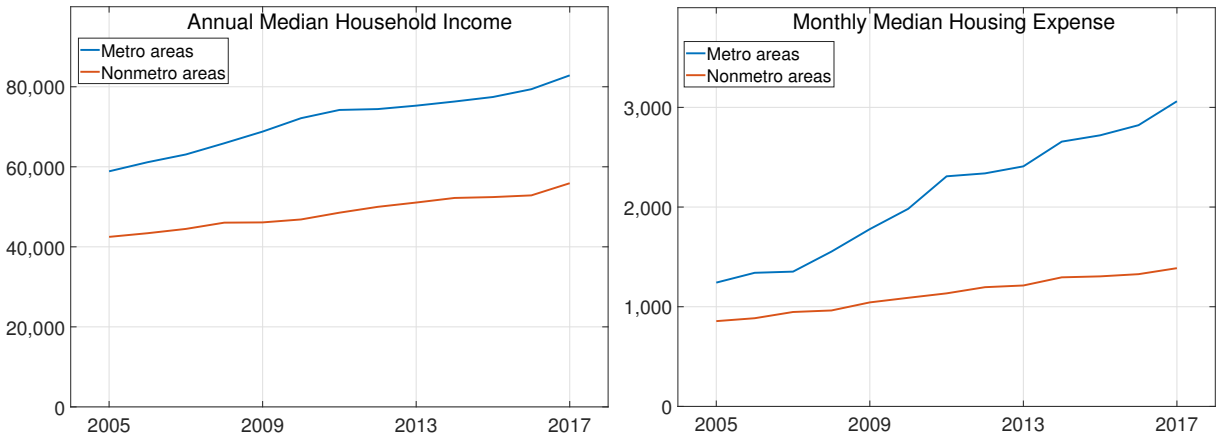
that benefit from the housing expansion need to compensate local residents, especially those who are economically disadvantaged.

1.10 Acknowledgement

Chapter 1, in full, is currently being prepared for publication. Jiajun Lu. The dissertation author was the sole author of this chapter.

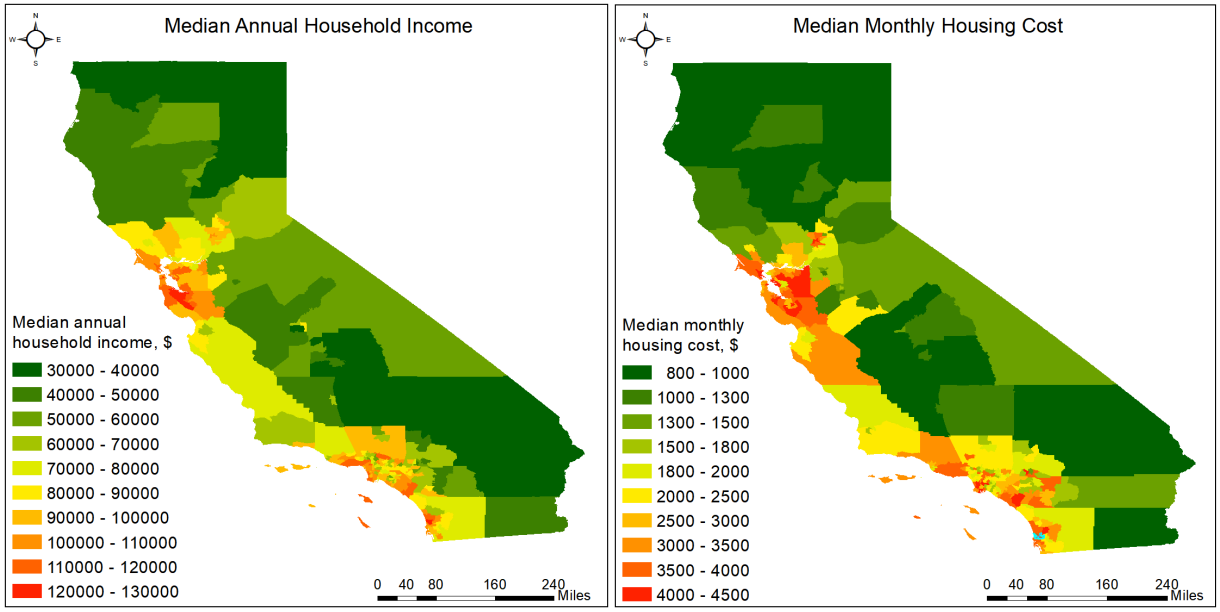
1.11 Appendix

1.11.1 Supplemental Table and Graph



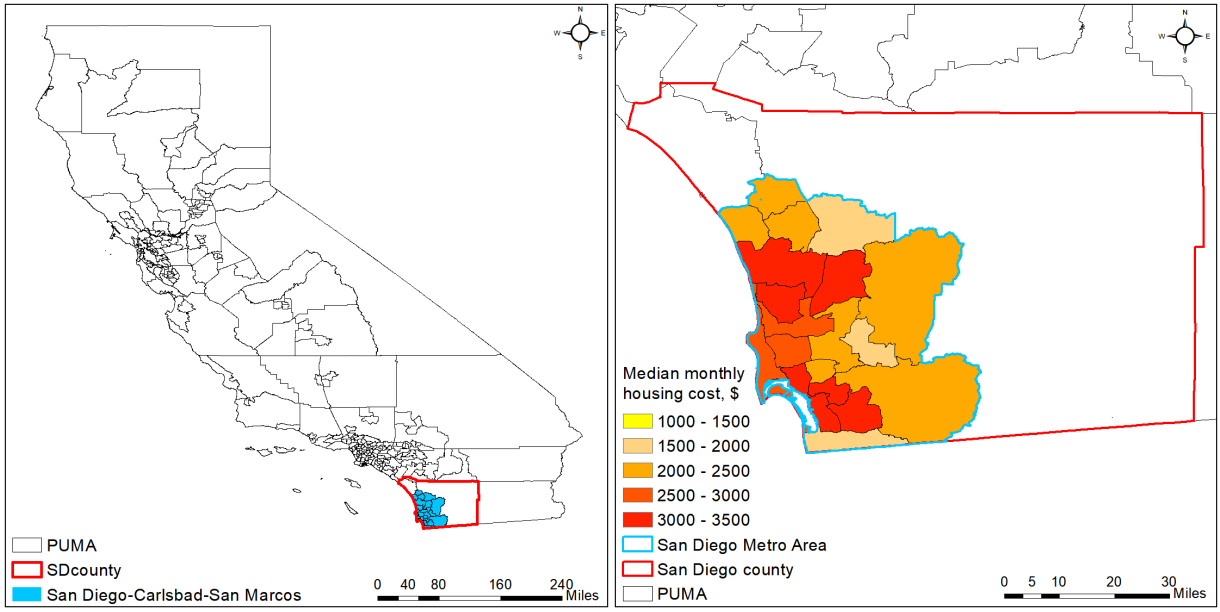
(a) Household Income (b) Housing Expense

Figure 1.22: Median Household Income and Housing Expense Trends between Metro and Non-metro Areas in California in 2005-2017



(a) Household Income (b) Housing Expense

Figure 1.23: Household Income and Housing Expenses across Geographical Areas in California



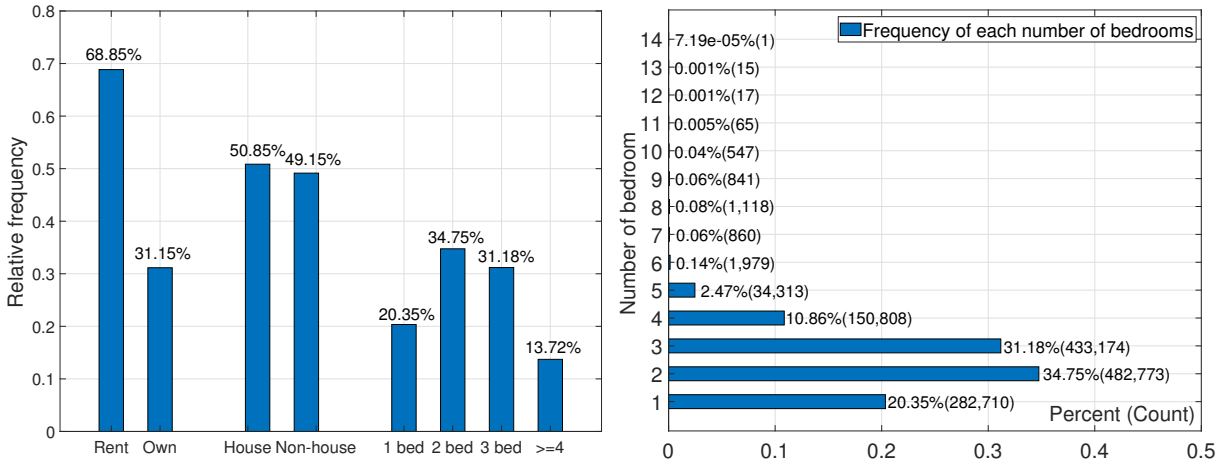
(a) Geographical Boundary in California (b) San Diego County

Figure 1.24: Delineation of PUMAs, Metro, and Nonmetro Areas in California and Median Annual Housing Costs across 20 PUMAs in San Diego County

Table 1.11: Summary Statistics of Wage Incomes by Occupations

Description of SOC Major Group	Average wage income	Percent (%)
Management	112,378.42	11.90
Business and financial operations	97,950,23	5.23
Computer and mathematicas	87,856.42	2.83
Architecture and engineering	85,342.32	2.36
Life, physical, and social science	73,964.43	1.03
Community and social services	63,847.38	1.84
Legal occupations	126,322.33	1.34
Education, training, and library	76,618.85	6.26
Arts, design, entertainment, sports	72,732.73	1.95
Healthcare practitioner and technician	91,969.34	5.41
Healthcare support	42,149.82	2.10
Protective service	50,310.38	2.36
Food preparation and serving	34,796.49	3.28
Building and grounds cleaning	41,786.03	3.57
Personal care and service	45,318.51	2.89
Sales and related occupations	55,476.81	10.12
Office and administrative support	53,626.79	12.96
Farming, fishing, and forestry	44,190.49	0.72
Construction and extraction	51,836.57	5.15
Installation, maintenance, and repair	48,420.18	3.74
Production occupations	48,478.47	6.49
Transportation and material moving	48,099.54	6.19
All employed	63,999.18	100

Note: The average wage incomes are measured in 2017 U.S. dollars. The occupational classification refers to Standard Occupational Classification (SOC) designed by the U.S. Census Bureau. See: http://www.census.gov/people/io/methodology/



(a) Summary Statisitics of Housing Choice (b) Number of Bedrooms

Figure 1.25: Sampling Distributions of Housing Choices

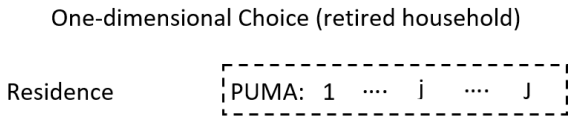


Figure 1.26: Residential Locational Choice by a Retired Household

1.11.2 Heterogeneous Housing Preferences

An illustrative model is proposed to conceptualize how preference heterogeneity in housing plays a role in the locational decision and forms a residential sorting. Assume a household gains utility from housing service (H) and non-housing composite good (N) and chooses to settle down between two residential areas, an expensive metropolitan area and a non-metro area, varying in both housing and labor markets. All households have the same preference for non-housing good but heterogeneous preferences for housing service. It is formulated by a classic Cobb-Douglas utility function, $U(H,N) = H^{\alpha}N^{\beta}$, with various α and the same β. Specifically, high-type households, who have a higher level of α, are willing to spend a larger fraction of income on housing service.

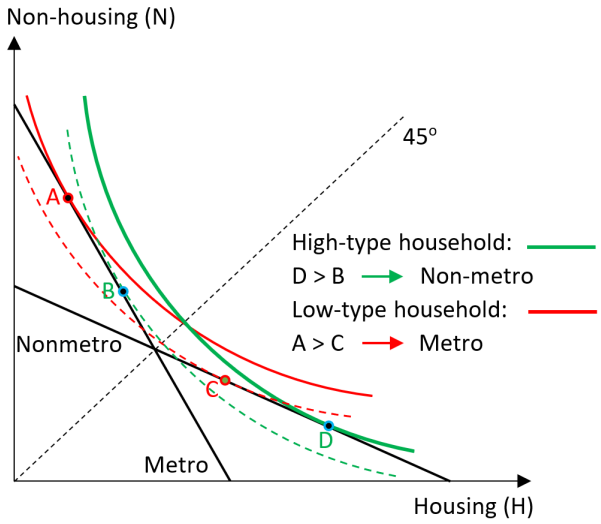


Figure 1.27: Residential Sorting between Metro and Non-metro Areas for Two Types of Households

Each economically rational household would make the simultaneous choice of a location and an optimal bundle of housing service and non-housing good. Figure 1.27 illustrates how the utility-maximizing choices are influenced by preferences for housing service. Compared to non-metro area, the metropolitan area corresponds to a steeper budget line starting with a higher point due to the wage premium and less affordable housing. All households have identical

demographics, thus having the same expected income and sharing the budget constraints in each area. For high-type households (green indifference curves), the optimal consumption bundle D in the non-metro area is preferred to B in the metropolitan area, making the non-metro area more desirable. In contrast, low-type households (red indifference curves) prefer bundle A over C and would move to the metropolitan area. The constrained optimization in each location is formalized as follows:

$$\begin{cases} \underset{H,N}{\text{Max}}\ U(H,N) = H^{\alpha}N^{\beta};\ \alpha,\beta \in (0,1),\ \text{s.t. } p_h H + p_n N \leq I \\ \text{Demand: } H^* = \frac{\alpha I}{(\alpha+\beta)p_h},\ N^* = \frac{\beta I}{(\alpha+\beta)p_n} \\ \text{Indirect utility: } V = U(H^*,N^*) = \left(\frac{\alpha I}{(\alpha+\beta)p_h}\right)^{\alpha}\left(\frac{\beta I}{(\alpha+\beta)p_n}\right)^{\beta} \end{cases}, \tag{1.14}$$

where p_h and p_n denote the unit prices of local housing service (H) and non-housing good (N), respectively, and I represents the household income. Given the budget constraint and specific preferences, households can predict the maximum utility from the optimal consumption bundle (H^*,N^*) in an alternative location. By comparing all maximum utility levels across local residential areas, households select the location with the highest utility to live. Then, I take the natural log of the indirect utility function in the equation (1.14) to make it linearly additive in preference and price parameters:

$$\begin{aligned} \text{In}V =& \alpha[\text{In}\alpha + \text{In}I - \text{In}(\alpha+\beta) - \text{In}p_h] + \beta[\text{In}\beta + \text{In}I - \text{In}(\alpha+\beta) - \text{In}p_n] \\ =& (\alpha+\beta)\text{In}I - \alpha\text{In}p_h - \beta\text{In}p_n - [\alpha\text{In}\alpha + \beta\text{In}\beta + (\alpha+\beta)\text{In}(\alpha+\beta)] \end{aligned}. \tag{1.15}$$

From the equation (1.15), it is shown that, when α gets higher, the housing price elasticity on the indirect utility becomes larger, which implies that high-type households are more sensitive to the change in housing price than low-type households. The preference heterogeneity in housing service can be identified by estimating distinctive preference parameters, α, across demographical groups. This explicit expression 1.15 sheds light on how to investigate the preference heterogeneity in housing service.

1.11.3 Simulation of Housing Preference Parameters

This section introduces how to simulate the household-specific housing preference parameters. Given the estimation results in Table 1.8, the housing preference parameter, β_1 is estimated to vary across households and follow the normal distribution, $\beta_1 \sim N(\mu,\ \sigma^2)$. Conditional on the household location choice, h_{ijk}, and observable household and locational attributes, Z_{ijk}, the conditional probability density function of the housing preference for household i, β_{i1}, can be derived using the Bayes rule as follows (Revelt and Train, 2000):

$$f(\beta_{i1}|h_{ijk},Z_{ijk},\mu,\sigma^2) = \frac{P(h_{ijk}|Z_{ijk},\beta_1)f(\beta_1|\mu,\sigma^2)}{P(h_{ijk}|Z_{ijk},\mu,\sigma^2)} \tag{1.16}$$

where $f(\beta_1|\mu,\sigma^2)$ is the overall distribution of the random housing preference parameter. h_{ijk} equals 1 if the residence j and workplace k are chosen by household i. Then, the household-specific mean of the parameter, $beta_{i1}$, becomes:

$$E(\beta_{i1}|h_{ijk},Z_{ijk},\mu,\sigma^2) = \int \beta_{i1} f(\beta_{i1}|h_{ijk},Z_{ijk},\mu,\sigma^2) d\beta_{i1}. \tag{1.17}$$

Intuitively, the expected value of β_{i1} can be thought of as the conditional mean of the coefficient distribution for the subsample who have the identical household demographics and make the same locational choice. In practice, the conditional expectation can be approximated using simulation as follows (Revelt and Train, 2000):

$$\widehat{\beta_{i1}} = \frac{\frac{1}{R}\sum_{r=1}^{R} \beta_{i1}^{[r]} \prod_{jk} \left[\frac{\exp(H_{ij}\beta_{1i}^{[r]} + Z'_{ijk}\theta)}{\sum_{jk}\exp(H_{ij}\beta_{1i}^{[r]} + Z'_{ijk}\theta)} \right]^{h_{ijk}}}{\frac{1}{R}\sum_{r=1}^{R} \prod_{jk} \left[\frac{\exp(H_{ij}\beta_{1i}^{[r]} + Z'_{ijk}\theta)}{\sum_{jk}\exp(H_{ij}\beta_{1i}^{[r]} + Z'_{ijk}\theta)} \right]^{h_{ijk}}}, \tag{1.18}$$

where $\beta_{i1}^{[r]}$ is the r-th draw for household i from the estimated distribution of β_{i1}.[74] All other variables are defined the same way as the equation (1.4).

[74]This paper draws 100 numbers for each household to calculate the household-specific coefficients on housing.

Chapter 2

Household Residential Location Choice in Retirement: The Role of Climate Amenities

2.1 Introduction

Locational choices among the retired population have long been a focal point in the analysis of social welfare. The large-scale migration resulting from millions of household relocations for retirement life can have a long-term impact on local population composition and thus become an important demographic and social phenomenon. During the past decade in the United States, migration contributed substantially to the aging of Florida and Arizona. These retirement migrations are expected to become even more salient as the population of baby boom generation continues to age. The members who were born during 1946-1964, the baby boom period, will pass the retirement age of 65 in 2020-2030, which makes the pool of potential retirement migrants reach a peak in the near future. The growing public attention on the upcoming retirement migration motivates the study of residential location choices of retired households in

this paper.

In addition to living costs and income, a household locational choice depends on many location-specific amenities. Among them, climate amenities can play an important role in choosing a residential location for retired households, given large geographical variations across climate regions in the United States. On account of large impacts on comfort, daily outdoor activity, and health, e.g., mortality risk, local climate amenities affect the desirability of different locations and the quality of life (Deschênes and Greenstone, 2011; Barreca et al., 2015). Households always prefer to retire in a place with favorable climate amenities, *ceteris paribus*. Thus, given their considerable influence on locational choices and household welfare, this paper conceptualizes the migration decision-making process driven by climate amenities and seeks to estimate dollar values retired household place on climate amenities. Past research has focused mainly on the middle-age or entire population, with relatively less attention paid to retired households who are more sensitive to climate amenities, especially temperature-related amenities (Albouy, 2016). Moreover, after retirement, household preferences for certain location-specific attributes, e.g., employment opportunities, can be different from younger working households (Chen and Rosenthal, 2008). However, no past research, to the best of my knowledge, has ever investigated retirement migration driven by climate amenities at the household level. In an attempt to fill the gap, this paper uses the newly released U.S. Census data to comprehensively analyze how locational decisions made by retired households are influenced by climate, with a focus on temperature-related amenities. In addition to temperature levels, retired elderly people may be sensitive to daily changes in the temperature. Given the rich information in available temperature data, this is the first paper that considers another temperature-related amenity, variability of temperature, and provides the relevant empirical evidence.

Unlike regular commodities that can be freely traded, climate amenities, as public goods, cannot be purchased separately and thus priced directly. Due to the lack of formal markets for them, estimating these values poses an econometric challenge since the evaluations of a friendly

climate would be entangled with many other local site characteristics. Given the development in estimation techniques, there are two main methodologies widely used in estimating climate amenities. The first is a hedonic pricing model assuming that decision-makers trade-off among economic variables, e.g., living cost, and other locational attributes, including climate amenities in choosing alternative locations (Rosen, 1974; Malpezzi, 2002). These estimated hedonic values are the empirical results as a consequence of a market equilibrium based on revealed preferences. However, these equilibrium outcomes obtained from the reduced-form hedonic models usually do not account for potential migration costs, leading to biased estimates of preference parameters in modeling locational choices. Therefore, a hedonic pricing model is unable to accurately measure climate values if the influence of migration costs is significant. This shortcoming motivates the use of the second approach, discrete choice model, that internalizes the high moving cost in a utility-consistent setting (McFadden, 1973). Under a utility maximization framework, the discrete choice modeling describes the choice behavior of retired households and yields unbiased estimates on values of climate amenities.

To account for potential preference heterogeneity in climate amenities, I construct a random coefficient logit model to value all primary climate amenities and examine how the marginal willingness to pay (MWTP) for temperature-related amenities varies with socio-demographic groups and residential location. It is found that, for a retired household, the average MWTP for a cooler summer by a 1°C is around $1,209, while the MWTP for a warmer winter by a 1°C is $1,114. Other than the mean temperatures, this paper reports the MWTP for a less variable temperature that older people may also favor and finds that they are willing to pay $486 for a 1°C drop in the average difference between daily maximum and minimum temperatures. This paper contributes to existing literature by further going deep down to the demographic categorization and examines the variations by age groups, household income levels, health status. It is found that older and wealthier retired households with a disability, *ceteris paribus*, have a higher MWTP

for preferred temperature amenities.[1] In addition to heterogeneous preferences by demographic attributes, I explore the geographical variations and find that households favoring the preferred temperatures more than the average live in places with a more friendly climate. As part of the estimation results, this paper updates the estimates of economic values for quality of life (QOL) in the U.S. metropolitan statistical areas (MSA) (Albouy et al., 2016). It shows that MSAs located on the east and west coasts have a better life quality, due mainly to a friendly climate and improved urban facilities, and the difference in the quality of life between two MSAs can be as large as $3,800, as measured in a dollar value.

Apart from the current generation, climate amenities can have a substantial long-term impact on the future retired population. The projected global climate change, however, can have an ambiguous impact on the desirability of local climate amenities in the United States. Due to rising temperatures driven by global warming, households suffer from hotter summers but benefit from milder winters. The changes in temperatures also depend on where households are located. Given the estimated value of temperature amenities and future temperature projections, I compute the value of projected changes in temperature amenities in 2050 and 2100 and find that households on average are willing to pay nearly 3.3% of their annual retirement income to avoid the future climate scenarios, conditional on current locations. Moreover, I simulate the location decisions of numerous retired households in new climate amenities and analyze the further residential sorting driven by the changing local climate. The simulation results forecast that hotter summers would overwhelm the warmer winters for future retired population and cause an overall northbound migration from the South climate region. Valuing future climate amenities not only advances the understanding of how climate affects social welfare but also the potential large-scale migration among the retired population. Ultimately, these findings have profound implications for local urban planning, e.g., urban facilities for retired population, in response to

[1] Compared to other studies that only roughly separate the entire population by a single cutoff of the age and the decision of relocation, this paper examines variations by other demographic attributes in more detail (Sinha et al., 2018).

changing demographic compositions through the long-term residential sorting.

The remainder of this paper is organized as follows. Section 2 presents a brief review of relevant literature. The econometric framework of a household locational choice model and empirical strategies are contained in section 3. The data used for empirical study are discussed in section 4. Section 5 presents the empirical results. In section 6, I value the projected changes in temperatures and simulate the residential sorting in response to climate changes. The paper concludes with a summary of key findings and policy implications in section 7.

2.2 Literature Review

A household location choice model regularly characterizes the selection of residential location by weighting site characteristics of each location for an economically rational household (McFadden, 1978). Typically, it assumes that a household utility is comprised of locational attributes, coupled with unobserved idiosyncratic errors, and estimates the hedonic value of some attribute through housing price or wage differentials across alternative locations. Previous studies have mainly focused on a couple of apparent motivations that affect household moving and location decisions. A large number of factors with some predictive power have been examined under the framework of random utility maximization (RUM) theory. Existing research has found that residential location choices can be motivated primarily by employment opportunity (Greenwood et al., 1991), education resources (Benabou, 1993), and transportation service (Anas, 1982).

In addition to urban facilities, the last few years have seen an increasing focus on natural amenities, especially climate amenities, in modeling residential locational choices. It is since, when living standard advances rapidly, climate amenities become much more prominent in evaluating the quality of life. A substantial body of literature, using a hedonic pricing analysis, has shown that households respond to climatic differences and value the favorable climate

amenities (Rehdanz, 2006; Butsic et al., 2011). However, some authors point out that high migration costs can introduce stickiness in a location choice and thus bias the estimates of climate values in a hedonic price model (Bayer and Timmins, 2007). In an attempt to incorporate the moving cost when estimating climate values, many studies have overcome the challenge and established a linkage between climate and household location choice. Poston et al. (2009) find that extreme temperatures have a large impact on migration flows and confirm that more friendly climates are positively correlated with in-migration rates. By constructing an inter-metropolitan residential location choice model, Plantinga et al. (2013) provide some estimates on the values of favorable changes in mean January and July temperatures. Sinha and Cropper (2013), using the U.S. census data, also characterize household locational choices among metro areas and provide empirical results about the influence of climate amenities on the desirability of alternative locations.

In a separate strand of literature, the driving force and motives of retirement migration have been fully explored, and many papers have analyzed how retirees make location decisions (Duncombe et al., 2001). The rationale behind these analyses is that, conditional on affordable moving costs, retired households vote with their feet and choose the utility-maximizing location by evaluating all locational attributes, including, among other things, climate amenities. As for socioeconomic characteristics, retired households are found to value low tax rates, crime rates, and low living costs, on the one hand (Longino, 1995). On the other hand, older retired persons have been shown to be attracted to locations with favorable natural amenities, such as sunny climates (Conway and Houtenville, 1998), access to coast (Bures, 1997), water area (Schneider and Green, 1992), and public parks (Duncombe et al., 2000). These findings provide guidance for choosing the main locational attributes that have a large influence on retired household utility in this paper.

The above-mentioned literature in the residential location choice model has been facilitated by the improved econometric techniques in discrete choice modeling. Over the past decades, the

modeling methods of residential household location choice have been largely developed (Hensher et al., 2005) and thus substantially contribute to the proliferation of researches in this area. Since the seminal paper by McFadden (1973), the multinomial logit (MNL) model has been the most common approach to modeling home location choice, due to a closed-form probability formula. The MNL model imposes an assumption of independence of irrelevant alternatives (IIA) among alternatives, which makes cross-elasticity across each pair of choice alternatives equivalent. Yet, the IIA assumption is violated if households perceive some destination alternatives as closer substitutes (Daly and Zachary, 1978). To address the issue of IIA constraint, a nested logit model (NL) allows alternatives to be grouped in a manner with correlations within, though not between, nests (Ben-Akiva et al., 1985). Soon after, to reflect a more flexible substitution pattern among alternatives, some more advanced discrete choice models are developed, such as ordered generalized extreme value (OGEV) model (Small, 1987), cross-nested logit (CNL) (Vovsha, 1997), and paired combinatorial logit (PCL) model (Wen and Koppelman, 2001). Despite a complex substitution pattern among alternatives, none of the discrete choice models consider potential heterogeneous preferences for certain attributes. To accommodate random tastes, McFadden and Train (2000) propose a mixed logit model that allows random coefficients. This flexible model features a framework that captures an unrestricted substitution pattern and individual-specific preference for some locational attributes in the decision-making process. Since then, due to its appealing property, the mixed logit model has been widely adopted in modeling a discrete choice in different contexts, including types of recreational activity episodes (Bhat and Gossen, 2004), electricity supplier (Revelt and Train, 2000), and drivers' parking (Chaniotakis and Pel, 2015) and driving behavior (Behnood et al., 2016). However, few papers have sought to model residential location choice with preference heterogeneity in some site attributes (Mistiaen and Strand, 2000). To fill the gap and contribute to the existing literature in residential location choice, the empirical analysis in this paper employs a mixed logit model that allows the coefficients on climate amenities to vary by households. This state-of-the-art modeling method fully reveals

heterogeneous preferences for climate amenities and estimates household-specific hedonic values of these favorable climate amenities.

2.3 Household Locational Choice Model

2.3.1 Utility Function Specification

To value climate amenities and examine their impacts on the decision of where to retire, I model household residential location choices under a random utility framework. Following the seminal work by McFadden (1973), retired households are assumed to be utility maximizers who attain utility through the selection of a preferred Metropolitan Statistical Area (MSA) in the United States. A household i, facing all alternative locations in the choice set ($j \in \mathbf{J}$), selects the location j and obtains a certain level of utility, U_{ij}, if and only if this alternative yields the highest utility, i.e., $U_{ij} > U_{il}$, $\forall\, l \neq j$. U_{ij} is a stochastic variable that can be decomposed into a systematic utility, V_{ij}, and a random part, ε_{ij}. The systematic component, V_{ij}, is a function of all observable attributes of alternatives and household characteristics, while ε_{ij} captures heterogeneity in preferences that are unobserved. This paper assumes that the utility of a retired household is dependent upon retirement income, housing cost, expense on non-housing services, climate and other locational amenities of the chosen residence, and moving cost in the relocation. Specifically, the utility that household i receives when living in MSA j is given by:

$$U_{ij} = V_{ij} + \varepsilon_{ij} = \alpha(Y_i - H_{ij} - Q_{ij}) + \Pi_j \beta_i + \Gamma_j \lambda + MC_{ij} + \eta_j + \varepsilon_{ij}, \tag{2.1}$$

where Y_i is the total income household i can receive in retirement. H_{ij} represents the housing expense and Q_{ij} denotes the cost of other non-housing services. In the baseline model, the household utility is assumed to be linear in the Hicksian bundle, $Y_i - H_{ij} - Q_{ij}$. The constraint of linearity in the Hicksian bundle is imposed to simplify the computation of welfare measures.

I relax this assumption with some non-linear specifications to check the robustness, as shown in Table 2.10 in Appendix. Π_j is a vector of observed location-specific amenities whose values vary across households. Γ_j represents locational attributes for which households have the same preference. Going forward, MC_{ij} is the general moving cost of a relocation that involves both economic and psychic costs. η_j is a locational fixed effect at the MSA level that controls for all unobserved location-specific amenities. ε_{ij} is the error term that incorporates unobserved utility-related preference heterogeneity.

To further elaborate the specification, some utility determinants are specified as follows. The total household retirement income, Y_i, is composed of negative incomes and assumed to be unrelated to locational choice.[2] The location-specific cost of other non-housing services, Q_{ij}, varies by household size. Housing expenditure, H_{ij}, is determined by the housing choice made by household i in MSA j. For simplicity, I predict the alternative housing expenditure with the assumption that a household consumes the same bundle of housing services. To test the validity of this assumption, I estimate the average number and standard deviations of some critical housing characteristics, e.g., the number of bedrooms and household tenure choice, across the MSAs for different demographic groups.[3] I find no significant variations in housing choices across MSAs in each group, which is in line with the conclusion by Sinha et al. (2018).[4] Thus, the alternative housing expense for each household i in MSA j is estimated and predicted based on the following hedonic housing equations for each MSA:

$$\ln H_{ij} = Z_i \beta^j + \varepsilon_{ij}, \forall j = 1, \ldots, J, \tag{2.2}$$

where H_{ij} is the annual housing cost of household i in MSA j. Z_i is the vector of housing choices

[2]The total household income involves all types of negative income, including retirement income, public assistance income, supplementary security income, and social security income.

[3]According to American Housing Survey, there exist small variations ($\leq 5\%$) in average bedroom size across metro areas. See: https://www.census.gov/programs-surveys/ahs.html

[4]Specifically, I use one sample t test for each housing characteristic and fail to reject the null hypothesis of the same housing choice for each demographic group.

and dwelling characteristics, and β^j are the MSA-specific coefficients for these housing attributes. Summary statistics of the estimation results of these hedonic housing equations are presented in Table 2.9 in Appendix.

When making a locational decision, moving cost is expected to deter a long-distance relocation and influence the utility of choosing each alternative location. To fully control for its impact, this paper adopts a more generalized form of moving cost involving both psychic cost of moving in various ranges (Davies et al., 2001) and economic cost.[5] Some papers set the birthplace of a householder as the origin of movement (Bayer et al., 2009; Fan et al., 2016). However, this assumption might not well apply to retired households, due to the fact that they are older and have less connection to the environment where they grew up. Instead, this paper takes the residence in which a household lived one year ago as the origin of movement. Specifically, based on geographic boundaries shown in Figure 2.1, the general moving cost is represented as follows:

$$MC_{ij} = \lambda_1 I_{ij}^{\text{Metro}} + \lambda_2 I_{ij}^{\text{State}} + \lambda_3 I_{ij}^{\text{Region}} + \lambda_4 d_{ij} + \lambda_5 d_{ij}^2, \tag{2.3}$$

where I_{ij} is a set of dummy variables reflecting the psychic cost in each range of movement. The dummy variables equal one if a household has to move out of certain MSA, state or region for an alternative location j. d_{ij} denotes the moving distance, and its quadratic form is used to proxy for the economic cost in the moving process.

2.3.2 Estimation Strategy and Choice Probability

To estimate heterogeneous preferences for climate amenities, I construct a mixed logit model that accommodates random coefficients (McFadden and Train, 2000). This paper mainly focuses on temperature amenities, and thus I allow the coefficients on three temperature-related attributes, i.e., $\beta_i = (\beta_i^{ST}, \beta_i^{WT}, \beta_i^{VT})$, to vary across retired households. β_i^{ST} and β_i^{WT} are the

[5]The psychic cost includes the loss of social network and familiarity with the surrounding environment, while the economic cost relates to what a household pays for a relocation.

coefficients on average summer and winter temperatures. The paper also incorporates another temperature-related amenity, the variability of temperature, as a climate amenity and its coefficient β_i^{VT} is assumed to be random as well. It is due to the fact that, generally, retired elderly people may be sensitive to daily changes in the temperature and its influence can vary across age groups.

To get around the potential bias from omitted locational and climate attributes in estimating the model, this paper adopts a two-stage estimation strategy (Murdock, 2006). The first stage is to estimate a mixed logit model where the systematic utility, V_{ij}, incorporates the Hicksian bundle, household-specific values for temperature-related amenities, general moving cost, and MSA fixed effects as below:

$$V_{ij} = \alpha(Y_i - \widehat{H_{ij}} - Q_{ij}) + \mathrm{WT}_j\beta_i^{WT} + \mathrm{ST}_j\beta_i^{ST} + \mathrm{VT}_j\beta_i^{VT} + MC_{ij} + \eta_j, \tag{2.4}$$

where $\widehat{H_{ij}}$ is the predicted housing cost. WT_j and ST_j are the average MSA-specific winter and summer temperatures. VT_j is the variability in the temperature, which is the average daily difference between the maximum and minimum temperatures in each MSA. All other variables are defined as the same as before. The three coefficients in β_i are assumed to be jointly normally distributed, with the mean vector μ and variance-covariance matrix Σ, i.e., $\beta_i \sim N(\mu, \Sigma)$. The matrix Σ is estimated in the first stage, while the means of β_i are restricted to be zeros. The mean coefficients of the three attributes, μ, can only be estimated in the second stage, since MSA fixed effects, η_j, technically absorb all the average influences of locational attributes. In the second stage, I regress the estimated MSA fixed effects on all climatic and locational attributes to estimate the mean coefficients of three temperature and other amenities as follows:

$$\widehat{\eta_j} = \Pi_j\mu + \Gamma_j\lambda + \omega_j, \tag{2.5}$$

where $\widehat{\eta_j}$ are the estimated MSA fixed effects. Π_j are the three temperature variables, i.e., WT_j, ST_j, and VT_j. Γ_j denote all other amenities and locational attributes for which households have

homogeneous preferences, λ. ω_j is the idiosyncratic error. This approach essentially treats the MSA fixed effect as a quality of life (QOL) index, which is equal to a weighted sum of climate amenities and other location-specific attributes (Albouy, 2016).

Assuming that the idiosyncratic errors, ε_{ij}, are independently and identically distributed with Type I extreme values, the probability of household i choosing MSA j is given by:

$$P_{ij} = \int \frac{\exp(\Pi_j \beta_i + Z_{ij}\theta)}{\sum_{j=1}^{J} \exp(\Pi_j \beta_i + Z_{ij}\theta)} f(\beta_i|\mu,\Sigma) d\beta_i, \tag{2.6}$$

where $f(\beta_i|\mu,\Sigma)$ is the density function of β_i that follows the multivariate normal distribution. Z_{ij} is the vector of all other variables for which households have the homogeneous preferences. Then, the parameters of equation 2.4 are estimated by maximizing the following simulated log-likelihood (*SLL*) function (Hole, 2007):

$$SLL = \sum_{i=1}^{N} \ln \left\{ \frac{1}{R} \sum_{r=1}^{R} \prod_{j=1}^{J} \left[\frac{\exp(\Pi_j \beta_i^{[r]} + Z_{ij}\theta)}{\sum_{j=1}^{J} \exp(\Pi_j \beta_i^{[r]} + Z_{ij}\theta)} \right]^{y_{ij}} \right\}, \tag{2.7}$$

where $\beta_i^{[r]}$ is the r-th draw from the joint normal distribution. R is the number of draws of random coefficients for each household. y_{ij} equals 1 if the household i selects alternative j and 0 otherwise.

2.4 Data

This section presents the data used in estimating the hedonic housing model and household locational choice model. This paper adopts a unique dataset, comprised of the U.S. census data and several other data sources.

2.4.1 Census Data

The main source of data used for the empirical analysis is the Public Use Microdata Sample (PUMS) from the American Community Survey (ACS).[6] It is a very detailed and comprehensive survey describing household socioeconomic and demographic characteristics. It covers around 1% of the entire population in the U.S at the household level on a yearly basis. Since climate amenities in each place typically change slowly within a couple of years or even decades, I select the newly released census data in 2017 to estimate location choice models.

2.4.2 Geography of the Choice Set

The study area of this paper is the entire continental United States, and I mainly focus on households residing in metropolitan areas. The study sample is chosen for two reasons. First, Hawaii, Alaska, and Puerto Rico are separate regions, and households have different preferences for climate amenities. This forms a household substitution pattern among alternative locations that are not comparable with the mainland United States. Secondly, the real moving cost between those regions and the continental U.S. cannot be well defined as that in the moving within the continental U.S., which may largely bias estimates on preference parameters in a locational choice model. Lastly, there exist rich data of urban amenities and locational attributes. The lowest level of identifiable location in PUMS is Public Use Microdata Area (PUMA), a statistical geographic area containing at least 100,000 people. Given that climate amenities do not vary significantly at a small geographic scale and most locational attributes are measured at the level of metro areas, this paper selects a metropolitan statistical area (MSA) as a choice unit by aggregating PUMAs into discrete MSAs. To map PUMA locations to the choice set of MSAs, I assign each PUMA to the MSA that overlaps its boundary. For the PUMA that belongs to several MSAs, I randomly assign the households into each MSA with the population-weighted probabilities. All 377 MSAs

[6] https://www.census.gov/programs-surveys/acs/

contain nearly 86 percent of the total U.S. population in 2017.[7]

Figure 2.1 illustrates the geography of the study area. Panel (a) shows that each retired household can choose among 377 MSAs to live. The color represents the percentage of the retired population in each MSA, which is the percent of retired households in the local population. It can be seen that the popular retirement spots are mainly concentrated in Southern California, Florida, and Texas. The metropolitan area, Villages in Florida, has the largest percentage of the retired population (56%). Panel (b) shows the geographic boundaries at various levels. Each MSA lies in a state that belongs to a division, and several divisions constitute a climate region. There are a total of 48 states plus District of Columbia, nine divisions, and four climate regions in the continental United States.[8] These boundaries divide the entire continent into several parts to estimate region-specific values of climate amenities. In addition, the geographical boundaries are delineated to calculate the psychic cost in the moving process.

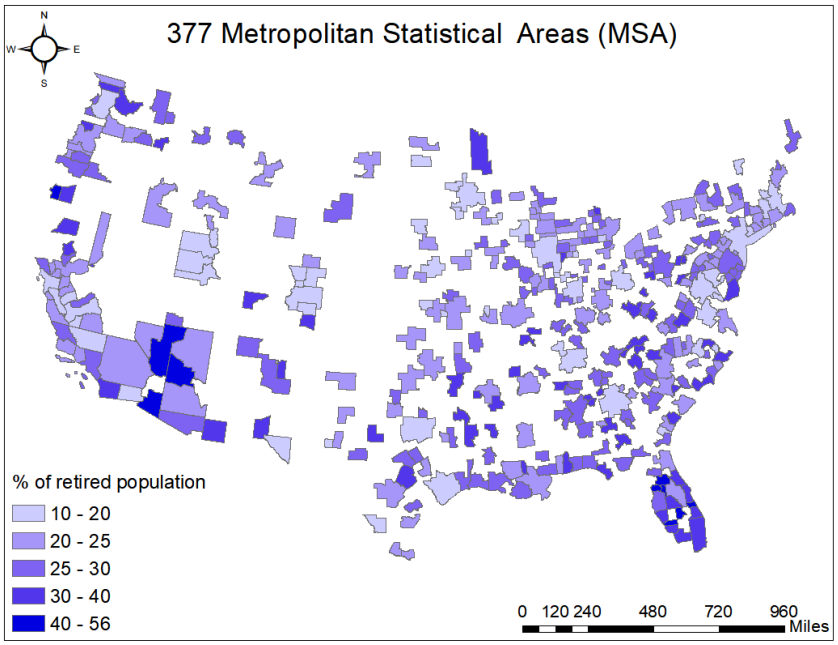
2.4.3 Sample Selection and Demographics

This paper mainly focuses on the influence of climate amenities on the life in retirement. I restrict the sample households to those who have retired as the decision-makers.[9] Households moving from areas other than the continental U.S. are dropped since they may have different preferences for certain attributes from local residents, leading to inconsistent estimates. Table 2.1 describes the characteristics of sample households. There are a total of 306,714 retired households, aged between 55-95 years old. I select only the households composed of a retired couple, a widowed or single retired person and drop multi-generational households (Lee and Painter, 2014). It is due to the fact that, if living with the next generation, locational choices

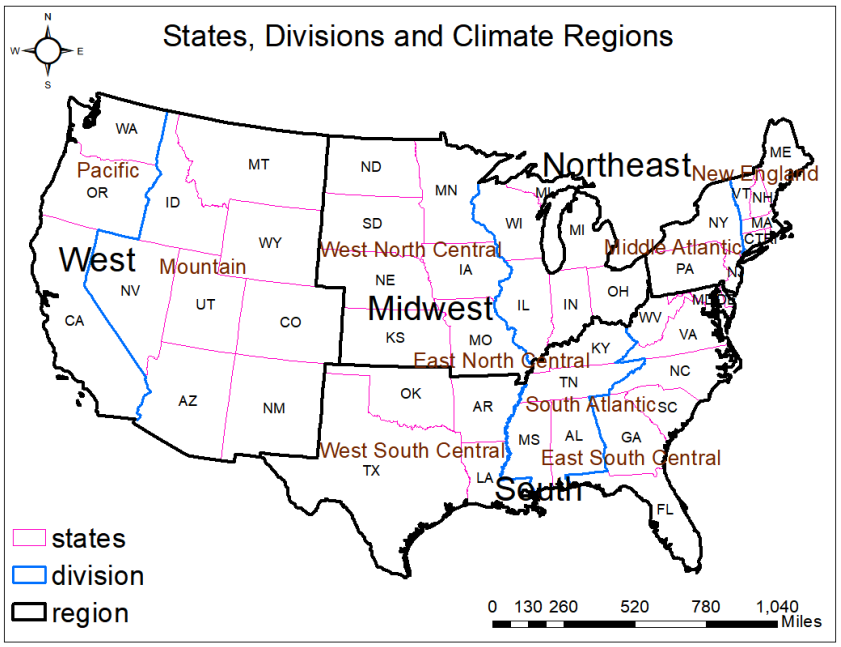
[7]The total of population estimate as of July 1 in those MSAs is 279.7 million, and the entire population in the U.S. in 2017 is 325.7 million.

[8]Northeast region consists of New England and Middle Atlantic divisions. The South Atlantic, West South Central, and East South Central divisions constitute the South region. The West North Central and East North Central comprise the Midwest region. The West region is composed of Mountain and Pacific divisions.

[9]Households are defined to be retired if they do not participate in the labor force and have no wage income.



(a) Geographic profile of MSAs



(b) States, divisions and climate regions

Figure 2.1: Geographic Profile of MSAs, States, Divisions, and Regions

of retired members can be constrained by other younger members, making the estimation of their preference parameters biased. It can be seen that most retired householders have a high school degree and over a third of them graduated from a college. Nearly 61% of respondents come from white households, while 17% of households have a disability. In terms of economic variables, it is shown that retired households on average have a household income of $42,002, lower than the entire population.[10] The average non-housing and housing expenses are almost 2/3 and 1/3 of their total retirement income. The data of livable non-housing expenses comes from Living Wage Calculator, a database that reports all expenses for a decent quality of life. It involves all basic needs, including food, childcare, health care, transportation, other necessities, and taxes.[11] The expenses vary by household size and are measured at the county level. To make living costs compatible with locational choices, they are converted into MSA level, weighted by population. Regarding the locational choices, it is seen that around 31% of retired households moved out of the previous metro area, and 8% moved out of a state. Only 5% of them relocated to a different climate region in 2017. The moving distances are the Euclidean distances between population-weighted centroids of two MSAs.[12] On average, each retired household moved to a residence 31.14 miles away from where they lived a year ago.

2.4.4 Housing Choice

As an important utility determinant, the housing service that a household attains has a large influence on locational choices. Table 2.2 presents the summary statistic of housing choices and property characteristics occupied by retired households. A retired household on average spends $16,051 per year ($1,337 per month) on housing, whose components depend on the tenure choice. The housing cost paid by renters includes the rent, insurance, and utility fees, while

[10]The average household income was $63,644 in the U.S. in 2017. I drop the sample whose household incomes are below $2,000 since their location choice can be very limited due to the budget constraint.

[11]http://livingwage.mit.edu/

[12]Due to large geographical variations in population density within each MSA, I calculate the geographic coordinates weighted by population density. https://www.census.gov/geo/reference/centersofpop.html

Table 2.1: Summary Statistics of Household Demographics and Locational Choices

Variable	Description	N	Mean	SD	Min	Max
Demographics						
Age	Age of household head	306,714	73.64	8.79	55	95
Household size	# of household members	306,714	1.63	0.55	1	2
High	High school graduate	306,714	0.90	0.30	0	1
College	College graduate	306,714	0.39	0.48	0	1
Female	Female household head	306,714	0.38	0.49	0	1
White	White household head	306,714	0.61	0.48	0	1
Dis	With a disability	306,714	0.17	0.47	0	1
White	White household head	306,714	0.61	0.48	0	1
Y	Household income	306,714	42,002.17	34,821.73	2,000	413,200
Q	Non-housing cost	306,714	25,073.09	8,083.76	2,939.60	35,484.87
H	Housing expenditure	306,714	16,051.13	12,091.21	504	160,704
Locational Choice						
I^{Metro}	Move out of a metro area	36,299[1]	0.31	0.40	0	1
I^{State}	Move out of a state	36,299	0.08	0.42	0	1
I^{Region}	Move out of a region	36,299	0.05	0.37	0	1
Moving distance	in miles	36,299	31.14	76.73	0	785.63

Note: The summary statistics are calculated by 306,714 retired households surveyed in 2017. [1]Among 306,714 retired households, 36,299 (12%) households moved in the previous year. Among All economic variables are measured in 2017 U.S. dollars.

the cost for homeowners involves property tax, homeowner association fee, insurance, utility fees, and, if applicable, home mortgage payment. Around 70% of households stay in their own dwellings, rather than renting. The percent of homeownership is slightly higher than that in the entire population, largely due to the higher wealth retired households have accumulated in the past.[13] Almost half of the dwelling units occupied by retired households are single-family houses and, on average, they need two bedrooms in each unit. Most of the dwelling units have complete facilities of kitchen, plumbing, running water, bathtub, and shower, which is important for the elderly. Nearly 3% of houses have a total lot size of more than ten acres. The average age of these housing units is 12 years.

Following the hedonic housing equation (2.2), I estimate the MSA-specific coefficients using the entire sample of 1,203,865 housing units in the study area.[14] The estimation results of these hedonic equations are presented in Table 2.9 in Appendix. It is shown that the means of most estimated coefficients are consistent with the conventional wisdom, even if they vary significantly across MSAs. Thus, housing costs need to be estimated separately across housing markets. Given the estimated coefficients and household housing choices, I predict the alternative housing expenditures for households have they lived in another place.

2.4.5 Climatic and Locational Attributes

In addition to the census data, climatic and locational attributes are attained to model household location choices. Among various climate amenities, temperature proves to be a primary concern for households (Schlenker and Roberts, 2009; Deschênes and Greenstone, 2011). This paper considers both mean and variability in temperature-related amenities. The mean

[13]The percent is around 62% of the entire population. It agrees with the life-cycle hypothesis that the elderly more likely have bought a real property, rather than rent, for the late time of residency (Green and Lee, 2016).

[14]I choose to estimate the housing equation over the entire sample since, from the statistical point of view, it is more accurate to estimate with all housing units, rather than only those occupied by retired households. The housing market is essentially exogenous for every single household. Moreover, I select only the private dwelling units and exclude some special residences, such as group quarters and public places.

Table 2.2: Summary Statistics of Housing Characteristics of Retired Households

Variable	Description	N	Mean	SD	Min	Max
Housing Choice						
H	Housing expenditure	306,714	16,051.13	12,091.21	504	160,704
Own	Owner-occupied unit	306,714	0.70	0.46	0	1
Property Characteristics						
house	Single-family house	306,714	0.51	0.49	0	1
bed	# of bedrooms	306,714	1.98	1.05	1	7
kit	Complete kitchen	306,714	0.93	0.08	0	1
plm	Complete plumbing	306,714	0.97	0.05	0	1
rwat	Running water	306,714	0.98	0.04	0	1
bath	Bathtub or shower	306,714	0.99	0.04	0	1
acr10	House on ten or more acres	306,714	0.03	0.17	0	11
year	Age of dwelling unit	306,714	12.37	11.05	0	74

Note: The summary statistics of housing choices and property characteristics are calculated over 306,714 retired households.

temperature in winter is measured over the three months from December to February, while the mean temperature in summer is the average from June to August.[15] The temperature variability is the average daily difference between the maximum and minimum temperatures. The three temperature-related attributes pick up most temperature impacts. Since climatic variables change slowly over decades and households mainly focus on the temperature in recent years, this paper computes the temperature variables over the past three years, i.e., 2015, 2016 and 2017. The climate data comes from GHCN-Daily, a dataset that contains daily maximum and minimum temperatures provided by the NOAA National Climatic Data Center of the United States (Durre et al., 2010).[16] The top panels (a) and (b) in Figure 2.2 illustrate the mean temperatures in summer and winter seasons measured in degrees Celsius, showing that there exist large variations in the average summer and winter temperatures across climate regions. The South region and southern Arizona in the West region experienced higher summer and winter temperatures than the rest of

[15]Existing literature also adopt the number of heating and cooling days in a year (Gyourko and Tracy, 1991) or the number of days in various temperature bins (Albouy, 2016).

[16]https://docs.opendata.aws/noaa-ghcn-pds/readme.html

the continental United States. The average summer temperature across 377 MSAs is 23.3°C with a standard deviation of 3.3°C. The average winter temperature is 4.7°C, and the standard deviation is 6.6°C.[17] The winter temperatures are, on average, more volatile than summer temperatures. Panel (c) displays the variability in the temperature across MSAs, and it can be observed that the West climate region has the largest average difference between daily maximum and minimum temperatures. It is partly because many MSAs are located in the desert climates in the West region. Panel (d) presents the spatial distribution of all 3,146 monitors covering the study areas.[18]

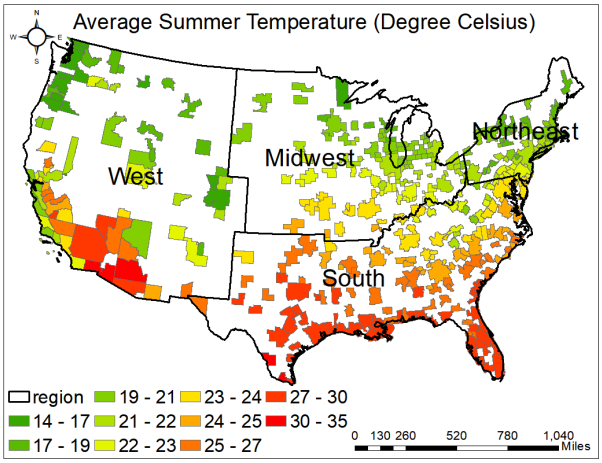
The dataset, GHCN-Daily, also reports many other climate attributes, including precipitation, snowfall, wind speed, particulate matter (PM2.5), and percent of possible sunshine. They are all included as utility determinants controlling for influences of these climate amenities. Table 2.3 summarizes these climate variables and shows the large variations in these climatic attributes across MSAs.

In addition to the common climate attributes, extreme weather and climate-related natural disasters, such as a hurricane in Florida and an earthquake in California, are likely to be considered by retired household in the locational decisions. However, due to the data limitation, it is impossible to collect all the information of many types of extreme weather, and they are barely comparable to each other in terms of the negative impacts. Moreover, retired households are likely to have both biased and heterogeneous perceptions of actual risks in these events, and the influence of extreme weather can be largely controlled for and predicted by climatic attributes, such as variations in temperature, wind, and precipitation.

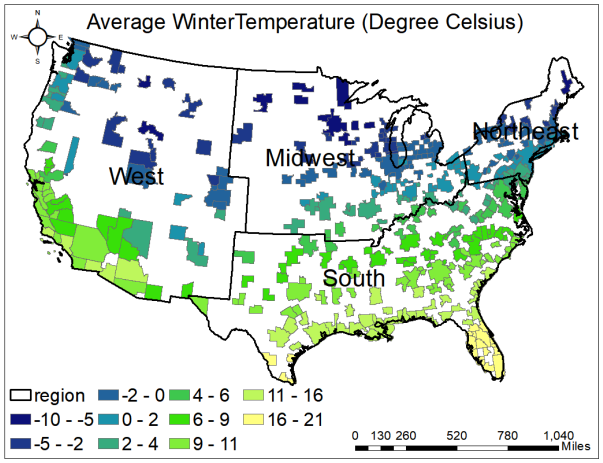
Given the data availability, many other non-climate locational attributes are also obtained from multiple sources and controlled for in estimating household location choices. I include as many location-specific amenities that retired household care about as possible in the model, as

[17]The winter and summer temperatures are highly correlated with the correlation coefficient of 0.87. The correlation coefficient is 0.23 between summer temperature and variability and 0.14 between winter temperature and variability.

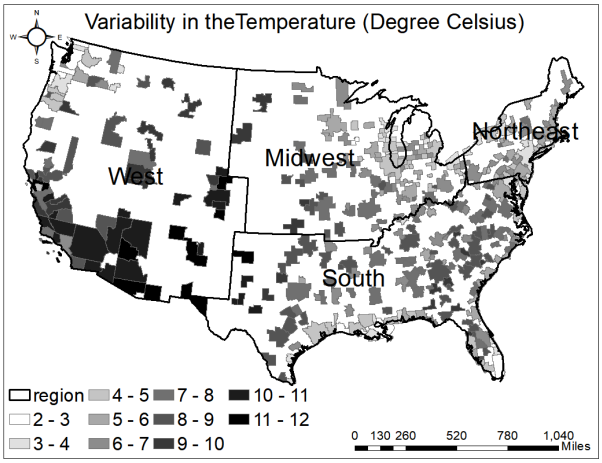
[18]Since the monitors are relatively evenly spread in each MSA, I give the readings of each monitor equal weights for the average climate attributes.



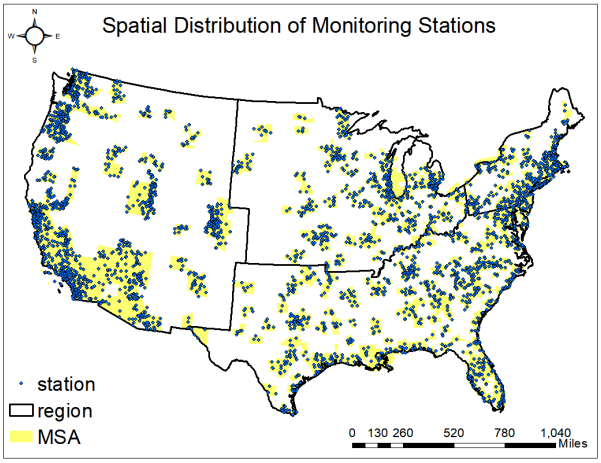
(a) Average summer temperature



(b) Average winter temperature



(c) Temperature variability



(d) Spatial distribution of stations

Figure 2.2: Daily Mean and Variability in Temperature

described in Table 2.3. The data of crime rate at each MSA comes from the FBI's Uniform Crime Reporting (UCR) Program.[19] The scores of education are obtained from WalletHub, a database that evaluates the average quality of educational system across MSAs in the United States.[20] The U.S. Department of Transportation reports transportation scores by MSAs in 2017.[21] The percents of the population with health insurance are estimated by the U.S. Census Bureau.[22] The data concerning park areas are provided by The Trust for Public Land, an organization reporting urban park statistics.[23] The values of non-climate locational attributes, in addition to climate amenities, are estimated with the equation 2.5 in the second stage of the locational choice model.

2.5 Empirical Results

This section presents the empirical results of the household locational choice model. I then investigate the preference heterogeneity in temperature amenities and the resulting residential sorting.

2.5.1 The Household Locational Choice Model

Given the choice set composed of 377 alternative locations for each of 306,714 retired households, I need to estimate the location choice model over 115,631,178 observations. To address the computational challenges, some papers adopt the sampling of alternatives that reduces the number of alternatives in the mixed logit model. It has been shown that a sampling strategy can theoretically produce consistent parameter estimates, but it loses some efficiency (Guevara and Ben-Akiva, 2013). By virtue of sufficient computing power in a computer server, the preference

[19] https://ucr.fbi.gov/crime-in-the-u.s/2017/
[20] https://wallethub.com/edu/most-and-least-educated-cities/6656/
[21] https://cms.dot.gov/transportation-health-tool/indicators
[22] https://www.census.gov/topics/health/health-insurance/data.html
[23] https://www.tpl.org

Table 2.3: Summary Statistics of Climatic and Locational Attributes

Variable	Description	N	Mean	SD	Min	Max
Climate attributes						
ST	Average summer temperature (°C)	377	23.33	3.31	14.11	34.42
WT	Average winter temperature (°C)	377	4.71	6.67	-10.27	20.92
VT	Variability in temperature (°C)	377	5.41	3.32	2.18	10.88
PRCP	Annual precipitation (inches)	377	39.72	14.81	9.51	60.12
SNOW	Annual snowfall (inches)	377	20.9	21.44	0.00	89.13
AWND	Average daily wind speed (miles/hour)	377	9.75	2.84	6.23	12.53
PSUN	Annual percent of sunshine	377	62.44	7.35	39.01	89.03
PM25	Mean PM2.5 (micrograms/m^2)	377	13.81	2.52	5.16	21.57
Locational attributes						
Elev	Average elevation (miles)	377	0.19	0.23	0.00	1.64
Disc	Distance to the coast (miles)	377	145.61	146.22	0.09	582.70
Pden	Population density (persons/$miles^2$)	377	246.98	258.61	7.46	2,316.02
Pwater	Percent of water area (%)	377	7.55	12.44	0.02	69.80
Trans	Transportation score	377	31.18	0.00	100.00	50.28
Crime	Crime rate per 100,000 inhabitants	377	450.98	208.14	5.87	1,300.07
Educ	Education score	377	50.23	25.32	0.00	100.00
Health	Health insurance coverage (%)	377	82.55	12.44	64.02	95.70
Park	Park area ($miles^2$)	377	201.10	382.31	1.13	797.61

Note: The summary statistics of climatic and locational attributes are measured at the MSA level in 2017. Temperature amenities are calculated over 2015-2017.

parameters in the household location choice model are estimated over the full choice set.[24] The model estimation is executed in PandasBiogeme, a free package for discrete choice modeling (Bierlaire, 2003).

Table 2.4 shows the estimation results of the household location choice model following the two-stage estimation strategy. In the first stage, the mixed logit model restricts the mean of $(\beta_i^{ST}, \beta_i^{WT}, \beta_i^{VT})$ to be zeros and allows the coefficients on three temperature-related variables to be jointly normally distributed. Therefore, only the standard deviations and correlations of the three coefficients can be estimated. It is seen that the level of preference heterogeneity in the winter temperature is slightly higher than that in the summer temperature, while the variation of coefficients on temperature variability is smaller than that in the mean temperatures. As for correlation coefficients, the winter and summer temperature coefficients are negatively correlated (-0.91). It suggests that retired households who prefer a warmer winter also favor a cooler summer, while those favoring a colder winter can sustain a hotter summer. In addition, the negative correlation between summer temperature and temperature variability implies that households who prefer cooler summers are more sensitive to changes in the outdoor temperature. Similarly, the preferences for milder winters are positively correlated with temperature variability, showing that households who value warmer winters more than others would like less volatile temperatures more than the average. Other than the statistics of household-specific coefficients, the first-stage model estimation yields the coefficients on the Hicksian bundle and generalized moving cost. I calculate the marginal willingness to pay (MWTP) on moving cost, climate and other location-specific attributes by dividing the associated coefficients by the coefficient on the Hicksian bundle, i.e., $E(\text{MWTP}^k) = -\frac{E(\beta^k)}{\alpha}$. It is estimated that the psychic costs of moving out of the MSA, state, and region in which a household lived before are \$416, \$887, and \$1,683, respectively. In terms of an economic cost in the relocation, it is observed that there exists nonlinearity in the relationship between a moving distance and an economic cost and a household

[24] Specifically, the model is estimated in C5 instance in the Amazon server. It performs in 3.0 GHz Intel Xeon Platinum processors, offering 72 vCPU and 144 GiB of memory.

pays $1,313 for a movement on average.[25]

As the group of coefficients estimated in the first stage, the estimates on locational fixed effects at the MSA level can be considered as overall evaluations for the quality of life in the metro areas. Figure 2.3 shows the economic values for the MSAs, among which Laredo in Texas state has the lowest estimate and is taken as the reference group due to its extremely hot summers. The economic values for quality of life are the estimated WTP for the differences between the reference location and any particular MSA. It is shown that, generally, MSAs located on the east and west coasts have a better life quality. Generally, retired households favor some popular retirement spots in Southern California, Florida, and Maryland more than other places. Salinas in California is given the highest dollar value, i.e., $3,750, compared to the reference group, largely due to its friendly climate, coastal attractions, and relatively high-quality urban facilities.

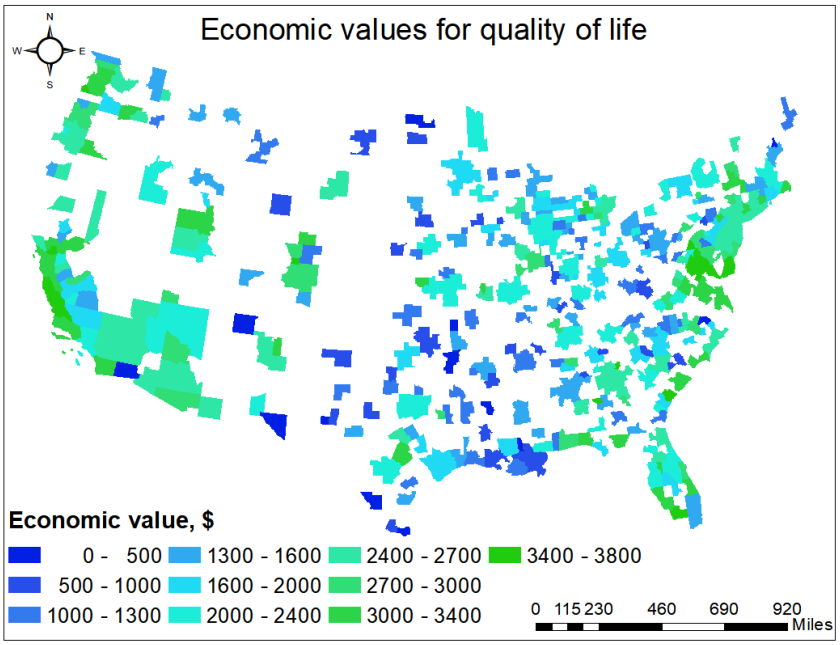


Figure 2.3: Economic Values for the Quality of Life across MSAs

In the second stage, the estimated MSA fixed effects are regressed on climatic and locational variables. The bottom panel in Table 2.4 reports both coefficients and MWTP on these

[25]The economic moving cost with an average moving distance is 7.9811*31.14+1.0987*31.14^2≈$1,313.94.

Table 2.4: Estimation Results of the Household Location Choice Model

Variables	Estimates (*util*)	Std Err	MWTP ($)	Std Err ($)
The first-stage estimation				
Dependent variable: deterministic utility, V_{ij}, in the equation (2.4)				
Std Dev of β_i^{ST}	0.0763***	0.0018		
Std Dev of β_i^{WT}	0.0867***	0.0027		
Std Dev of β_i^{VT}	0.0332***	0.0081		
Correlation coefficient				
$\rho(\beta_i^{ST}, \beta_i^{WT})$	-0.9111***	0.0134		
$\rho(\beta_i^{ST}, \beta_i^{VT})$	-0.4341***	0.0287		
$\rho(\beta_i^{WT}, \beta_i^{VT})$	0.3127***	0.0170		
Hicksian bundle	4.7621e-5***	1.1862e-5		
Moving out of metro	-0.0801***	0.0176	-1,683.18	369.83
Moving out of state	-0.0426***	0.0075	-886.76	156.11
Moving out of region	-0.0198***	0.0065	-416.09	136.60
Moving distance	-3.8706e-4***	8.7036e-6	-7.9811	0.1794
Moving distance squared	-5.2321e-5***	3.3006e-7	-1.0987	0.0069
# of observations=115,631,178, # of decision makers=306,714				
The second-stage estimation				
Dependent variable: the estimated MSA fixed effects, $\widehat{\eta}_j$, in the equation (2.5)				
Mean summer temperature	-0.0576**	0.0290	-1,209.97	611.09
Mean winter temperature	0.0531**	0.0250	1,114.09	525.51
Variability in temperature	-0.0231*	0.0128	-486.13	270.07
Annual precipitation	0.0052*	0.0028	109.09	60.60
Annual snowfall	-0.0189**	0.0094	-398.12	199.06
Average daily wind speed	-0.0008	0.0006	-18.12	13.93
Annual percent of sunshine	0.0047*	0.0026	98.24	54.57
Mean PM2.5	-0.1528**	0.0764	-3,210.01	1,605.01
Average elevation	-0.0002	0.0002	-5.12	5.12
Distance to the coast	-0.0013**	0.0006	-28.23	14.11
Population density	0.0043	0.0053	91.42	114.27
Percent of water area	0.0375*	0.0208	789.21	438.45
Transportation score	0.0391**	0.0195	821.13	410.56
Crime rate	-0.0052**	0.0026	-109.27	54.63
Education score	0.0014	0.0017	29.56	36.95
Health insurance coverage	0.0104**	0.0051	218.09	108.50
Park area	0.0230***	0.0058	483.87	124.07
# of observations=377, Adjusted R^2=0.4769				

Notes: The marginal willingness to pay (MWTP) is calculated by normalizing the coefficients on Hicksian bundle, measured in 2017 dollars. Robust standard errors are reported in the right columns, *** p<0.01, ** p<0.05, * p<0.1.

local attributes. It shows that retired households view higher winter temperatures and lower summer temperatures as preferred climate amenities. On average, they are willing to pay $1,209 for a 1°C decrease in average summer temperature. This large MWTP should be interpreted as the value of a cooler summer during the entire three months. On the other hand, cold winter is considered a disamenity and retired households are willing to pay $1,114 for 1°C increase in the winter temperature. The MWTP of temperature amenities are nearly 2.9% and 2.7% of average annual household incomes among the retired population. By comparison, the percents are much higher than those in the classic and well-cited paper by Albouy et al. (2016). They find that the WTPs to reduce an additional heating degree and cooling degree are 0.8% and 1.9% of income in the entire population. The considerable difference is largely due to the fact that residents become more sensitive to temperature and thus value them more highly when getting older.[26] It can also be the case that retired households care less about other attributes, like employment opportunities and school education for the next generation, and thus put a higher weight on climatic amenities than younger households in a locational decision (Lee, 2018). Moreover, the mean MWTP for a decrease of 1°C in the daily difference between the maximum and minimum temperatures is $486, nearly a half as much as that for a change of 1°C in the summer and winter temperature. This new evidence suggests that retired households also value daily temperature change. Apart from average temperatures in summer and winter seasons, the second-stage estimation yields the coefficients and MWTPs for other attributes. The empirical results show that retired households view a higher level of precipitation, a larger percent of sunshine, proximity to the coast as valuable natural amenities, while a higher level of snowfall and air pollution, higher wind speed, and higher elevation are taken as disamenities. In terms of locational attributes, they prefer a higher population density, larger percent of water area, an advanced transportation facility, and lower local crime rate. As opposed to climatic attributes that are exogenous, some might be concerned that there exist general equilibrium effects of a climate-driven migration. It is due to the fact

[26] Another similar evidence comes from the study by Sinha et al. (2018). They estimate that the MWTP for summer and winter temperatures are $873 (1.4%) and $709 (1.1%), respectively.

that numerous locational choices made by retired households, in the aggregate, can influence the local locational attributes, such as the density of retired population, that, in return, have an impact on locational decisions. The potential endogeneity arising form the reverse causality can bias the estimation results. To address the concern, I extend the model by adding two more variables related to the local age distribution, the percent of the retired population and density of the retired population, in estimating the main mixed logit model. However, having controlled for the local population density, none of the two variables are statistically significant.[27] Furthermore, retired households also favor a better medical system and more parks in an MSA, while, as expected, they barely have any preference for local education quality. Among these site characteristics, the coefficients on air quality, convenient transport, coast amenity, public safety, health system, and park facilities are statistically significant, showing that retired households have stronger preferences for these attributes. The estimated preferences for some attributes conform to the features of life in retirement.

The baseline model with the utility specification 2.4 assumes that 1) retired households have identical preferences for marginal changes in economic variables and 2) moving cost depends on the moving distance and the place where you lived one year ago. Table 2.10 in Appendix reports the robustness check of the value of temperature amenities to the setting of economic components and moving cost. Specifically, the Model 2 relaxes the linearity in the Hicksian bundle and allows the coefficients on the retirement income net of housing cost and non-housing cost to be different. It is shown that the relaxation of linearity in the Hicksian bundle has no significant influence on the estimates of the coefficients in Model 2. Following the setting of moving cost with the birthplace of a householder as the origin (Bayer et al., 2009; Fan et al., 2016), Model 3 checks how the hedonic values for temperature amenities can be influenced by moving cost. The alternative moving cost becomes $MC_{ij} = \pi_1 I_{ij}^{\text{Metro}} + \pi_2 I_{ij}^{\text{State}} + \pi_3 I_{ij}^{\text{Region}}$, where

[27] Admittedly, some other urban facilities in the supply side might be endogenous, such as an increased supply level of housing for the retired population that influences local housing expenses. For instance, some retired households do not like to be surrounded by many retired households but prefer living with younger people.

the dummy variables, I_{ij}, equal one if choosing a place different from the birthplace in each range. It is observed that the magnitudes of MWTP for staying in the same birthplace are much lower than those for the same current residence and the estimated values of temperature amenities are also lower than those in the baseline model. The large differences show the importance of moving costs in a locational choice. Retired households value the recent residential area more highly than their birthplace since they have less connection to the environment of birthplace when they retire. Therefore, it is reasonable to keep the setup of the baseline model with a linear Hicksian bundle and moving cost determined by the previous living area for the rest of the empirical analysis.

2.5.2 Heterogeneous Preferences for Temperature Amenities

Among the retired households, values of temperature amenities can differ depending on demographic attributes and other unobserved preferences. The random coefficients on the temperature-related variables in the mixed logit model enable the exploration of preference heterogeneity in these amenities. Conditional on the household location choice, y_{ij}, and observable household and locational attributes, Z_{ij}, the conditional distribution of the random coefficients can be derived using the Bayes rule (Revelt and Train, 2000):

$$h(\beta|y_{ij},Z_{ij},\mu,\Sigma) = \frac{P(y_{ij}|Z_{ij},\beta)f(\beta|\mu,\Sigma)}{P(y_{ij}|Z_{ij},\mu,\Sigma)} \tag{2.8}$$

where $f(\beta|\mu,\Sigma)$ is the overall distribution of random parameters. y_{ij} equals one if alternative j is chosen by household i. Then, the household-specific means of the parameters become:

$$E(\beta_i|y_{ij},Z_{ij},\mu,\Sigma) = \int \beta_i h(\beta|y_{ij},Z_{ij},\mu,\Sigma)d\beta. \tag{2.9}$$

Intuitively, the expected β_i can be thought of as the conditional means of the coefficient distribution for the subsample who have the identical household demographics and make the same locational

choice. In practice, the conditional expectations can be approximated using simulation (Revelt and Train, 2000) as follows:

$$\widehat{\beta}_i = \frac{\frac{1}{R}\sum_{r=1}^{R}\beta_i^{[r]}\prod_{j=1}^{J}\left[\frac{\exp(\Pi_j\beta_i^{[r]}+Z_{ij}\theta)}{\sum_{j=1}^{J}\exp(\Pi_j\beta_i^{[r]}+Z_{ij}\theta)}\right]^{y_{ij}}}{\frac{1}{R}\sum_{r=1}^{R}\prod_{j=1}^{J}\left[\frac{\exp(\Pi_j\beta_i^{[r]}+Z_{ij}\theta)}{\sum_{j=1}^{J}\exp(\Pi_j\beta_i^{[r]}+Z_{ij}\theta)}\right]^{y_{ij}}}, \tag{2.10}$$

where $\beta_i^{[r]}$ is the r-th draw for household i from the estimated distribution of β. This paper takes 100 draws for each household to calculate the household-specific coefficients on temperature-related variables.

The MWTP for temperature amenities may be influenced by age and health status due to their close relations with the desire for a friendly climate in doing outdoor activities. Moreover, the budget constraint can also have an impact on how much a household is willing to pay for a preferred temperature. Table 2.5 reports the means of the conditional MWTP that are averaged across all households in each subgroup divided by age, income level, and health condition. It can be seen that older retired households do favor a friendly temperature more than the younger retired groups, confirming the stronger desire for warmer winters, cooler summers, and steady weather for outdoor activities. This trend lasts until households turn the 90s, partially because they stay longer in the rooms and thus no longer need the preferred outdoor temperature that much. In another aspect, household income has a significant influence on the MWTP for temperature amenities. This positive relationship between income level and MWTP reflects a higher cost richer households are willing to pay for climate amenities and related well-being. In terms of health status, the mean MWTP for preferable temperature amenities is higher for disabled households than those without a disability, due to the fact that, generally, a disability makes it harder to do daily outdoor activities in a bad climate.[28]

[28]The defined disability in the census data does not specify the type of physical features, which thus relates to an overall impact.

Table 2.5: Heterogeneous Preferences for Temperature Amenities by Demographic Groups

MWTP ($)	Mean summer temp	Mean winter temp	Variability in temp
Age group[a]			
55-65	-1,071.72	994.19	-387.31
66-75	-1,223.14	1,043.89	-498.34
76-85	-1,423.09	1,344.93	-548.26
86-95	-1,119.90	1,074.19	-423.53
Household income			
1%-25% ($2,000-$18,400)	-1,129.73	914.26	-246.13
25%-50% ($18,400-$32,300)	-1,285.06	1,181.75	-316.23
50%-75% ($32,300-$54,100)	-1,289.37	1,224.45	-436.34
75%-100% ($54,100-$413,200)	-1,310.93	1,243.09	-587.09
Health status			
With a disability	-1,313.42	1,244.29	-656.71
Without a disability	-1,167.56	1,002.24	-386.36
All	-1,209.97	1,114.09	-486.13

Notes: [a]The age group is categorized by the average age of household members.

2.5.3 Residential Sorting for Temperatures

Apart from the preference heterogeneity across demographic groups, there could exist a residential sorting based on the preferences for temperature-related amenities among retired households. Given the household locational choices, I calculate and average the household-specific coefficients on temperatures for each MSA, in order to examine how households in retirement sort across locations with their tastes for temperature amenities.

Figure 2.4 shows the relationship between temperature amenities and average household MWTP for them across the MSAs. Each dot represents an MSA, and they are categorized by climate regions and represented by various symbols and colors. As seen in Panel (a), there exists a strong negative correlation between MWTP for cooler summers and MSA summer temperatures. It indicates a residential sorting pattern across cities. Holding other factors equal, households with a higher MWTP for lower summer temperatures tend to retire in a cooler city in summer, while those with a lower MWTP have resided in hotter cities. The three salient metropolitan

areas on the top bottom corner are Phoenix, Yuma, and El Centro in Arizona and California. It is intuitive since these places located in the desert experience the highest temperatures during summer days with no precipitation. Retired households who choose to stay there thus have a higher ability to endure hot summers and do not have much incentive to move out, based on their revealed preferences. The Panel (b) illustrates the taste-based sorting for winter temperatures, and we can observe a positive relationship between winter temperature and its average MWTP across climate regions. Specifically, retired households residing in the South climate region with higher winter temperatures typically favor and thus are willing to pay more for warmer winters than those living in other climate regions. It largely explains why households who have a higher MWTP for warmer winters are more likely to overcome a moving cost and choose to live in the South region for retirement. On the contrary, many MSAs in the Midwest, Northeast and West regions (Northern part of the continental U.S.) feature lower winter temperatures, some of which are even below 0°C. These urban areas are mainly inhabited by retired households who are willing to pay less than the average for warmer winters. As to the daily variability in temperature presented in Panel (c), there exists a less obvious sorting pattern, and retired households barely have significantly different MWTPs for this temperature amenity across MSAs. Some MSAs in western regions, located in the top right corner, feature higher average differences in daily temperature. However, households who choose to retire there have a similar MWTP for temperature variability to other MSAs, suggesting that temperature variability is not the primary driver for climate-related migration.

Table 2.6 reports the temperature amenities and calculates MWTP conditional on the locational choices across climate regions. The MWTPs for temperature-related amenities are first averaged across all households in an MSA and then weighted by the MSA population to obtain region-level values. It can be seen that, due to the preference-based sorting across MSAs, the population-weighted MWTPs for summer and winter temperatures are higher than those without sorting. The average MWTPs for a warmer winter and cooler summer are higher in the South

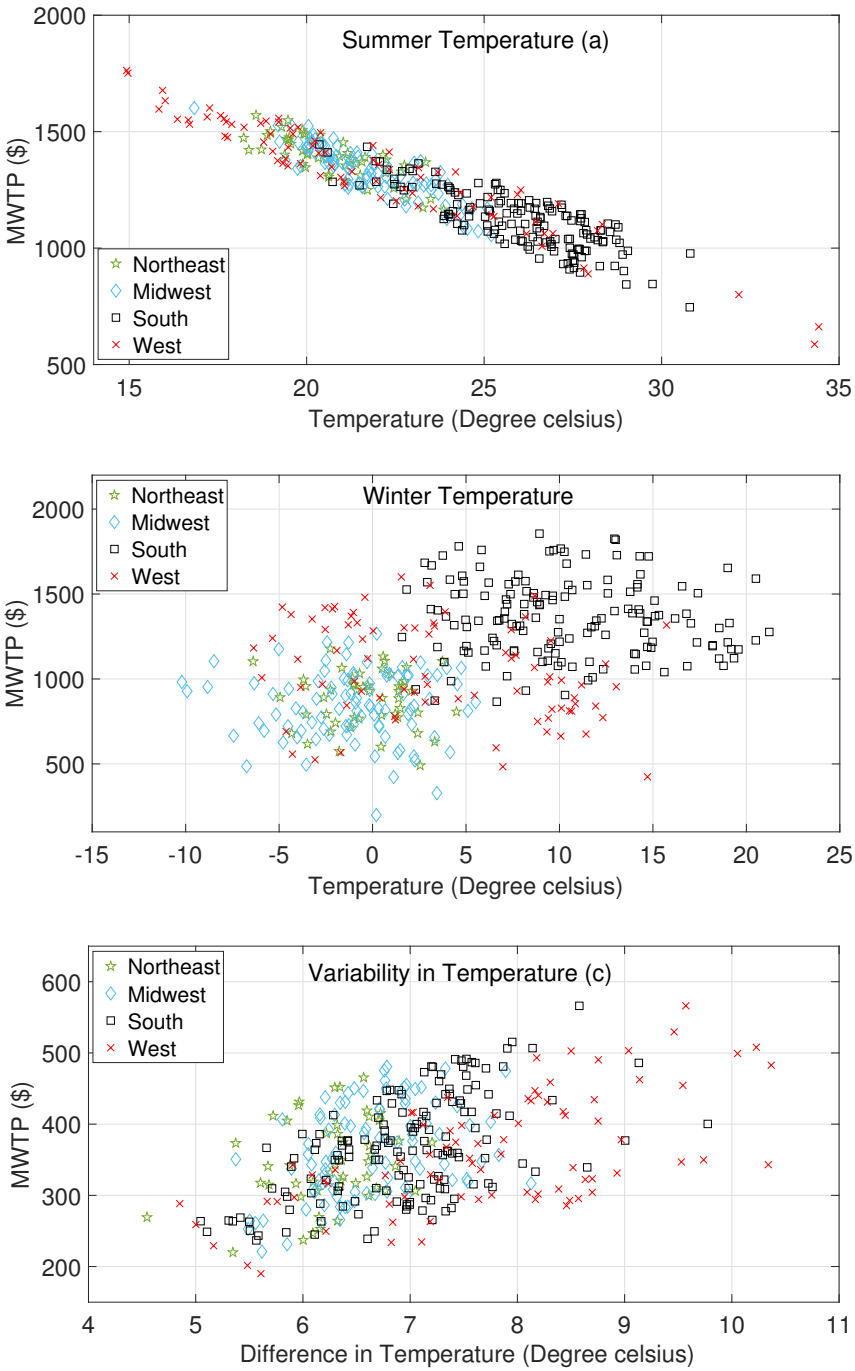


Figure 2.4: Residential Sorting across MSAs Based on Temperature Amenities

region than the rest, while the average MWTP for variability in temperature does not vary mainly by climate regions.

Table 2.6: Temperature Amenities and MWTP by Climate Regions

Region	Midwest	Northeast	West	South	All
Temperature amenities					
Summer temperature	21.3	17.8	27.8	30.2	23.8
Winter temperature	-3.5	-4.6	6.8	10.7	4.5
Variability in temperature	5.0	4.9	5.9	4.7	5.3
Marginal willingness to pay (MWTP)					
MWTP for summer temperature	-1,313.1	-1,210.9	-1,494.5	-1,565.1	-1,416.1
MWTP for winter temperature	1,041.5	1,144.3	1,342.0	1,412.0	1,294.3
MWTP for temperature variability	-474.1	-461.0	-459.3	-479.5	-471.9

Notes: The temperature-related variables and MWTPs are first averaged across all households in an MSA and then weighted by MSA populations to gain region-specific values.

In sum, there exist significant preference-based sorting patterns for both summer and winter temperatures, while the taste-based sorting does not exist for the variability in temperature. Retired households who favor the preferred temperatures more than the average live in places with a more friendly climate, while those who can endure a hot summer and cold winter retire in residences with less preferred climate amenities. The revealed preferences in the locational choices are used to solicit household MWTP for these climate amenities, showing that residences with preferred temperatures are indeed inhabited by retired households who are willing to pay more for these amenities. The MWTP for a better temperature in some MSAs can be four times ($2,000 vs. $500) as high as that in other MSAs, which implies retired households have very different valuations for climate amenities across the United States. Thus, taste-based sorting should not be ignored when estimating the aggregate value of climate amenities.

2.6 Values of Projected Temperature Changes

Using the estimated preference parameters for temperature amenities, this section provides the empirical estimates on values of future temperature changes for retired households.

2.6.1 Projections of Temperature Amenities

Many projections of future temperatures have been made by climate scientists under various climate scenarios and models. To value changes in the temperature amenities, I apply the most commonly used climate projection dataset, NASA Earth Exchange Global Daily Downscaled Projections (NEX-GDDP).[29] This NEX-GDDP dataset includes downscaled projections from the 21 models and scenarios for which daily scenarios are produced and distributed under the Coupled Model Inter-comparison Project Phase 5 (CMIP5) (Taylor et al., 2012). It presents the global climate projections of daily temperature over the periods from 1950 through 2100 at very small scales.[30] Specifically, I select two typical time points, the year 2050 and 2100, under the scenario of RCP45.[31] The average summer and winter temperatures are then calculated in these two years for each MSA. A total of 4,349 locations with projections spatially overlap with the MSAs. The values are averaged over all spots in each MSA. Figure 2.5 illustrates the projected changes in summer and winter temperatures under various situations. It can be seen that, in 2050 and 2100, both the summer and winter temperatures are projected to be higher, which implies that we will experience warmer winters and hotter summers this century. The temperature projections for the continental United States do not largely differ between the two periods. In the coming 2050, there will be an average increase of 3.1°C in summer temperature and 2.9°C in winter temperature, which are slightly lower than those in 2100 (3.6°C in summer and 3.3°C in winter).

[29] https://cds.nccs.nasa.gov/nex-gddp/

[30] The spatial resolution of the dataset is 0.25 degrees (-25 km x 25 km), and there are a total of 1,036,800 spots with projections across the globe.

[31] RCP is short for Representative Concentration Pathways. Another common climate scenario is RCP85. The two climate scenarios vary by the projected concentration of greenhouse gas emissions but with an only slight difference in projected temperatures (Karl et al., 2009).

The average temperature variability almost remains the same change over decades, and thus its calculated WTP is not reported. As for the variations across climate regions, summers in the future would become much higher in the South region, compared to the rest, while the West region will experience warmer winter than other climate regions. Therefore, the changing temperature amenities caused by long-lasting global warming in the 21st century can have location-specific impacts on retired households. I calculate the dollar values of these climate changes to evaluate its influence on daily life and welfare.

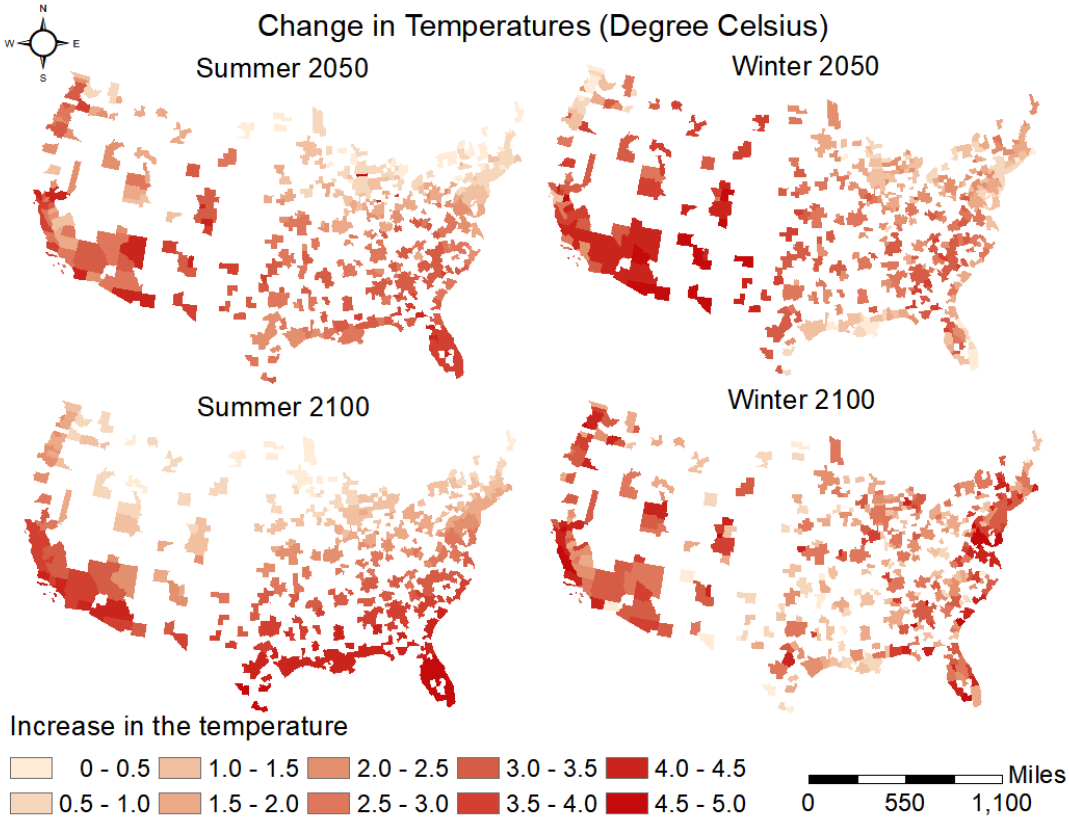


Figure 2.5: Δ°C in 2050 and 2100 across MSAs

2.6.2 WTP for Temperature Change with Current Locations

Given the coefficients on temperature amenities and projected change in the temperatures, I compute WTP by multiplying the MWTP for summer and winter temperatures in each MSA by

the size of the temperature change, conditional on household location choices. The population-weighted WTPs in each climate region and the entire U.S. are also computed for each climate scenario. Table 2.7 presents the changes in temperatures across regions and the WTP for each situation conditional on locational choices. The WTP equals the value of warmer winters, net of the disvalue of hotter summers. It is seen that, in terms of the entire U.S., the overall WTP for climate change is negative, implying that global warming will cause aggregate damage to the wellbeing of the retired population. A retired household on average is willing to pay $890 (nearly 2.1% of an annual retirement income) to avoid the climate scenario projected to occur in 2050 and $1,379 (3.3% of their annual income) in 2100. The empirical estimates show that retired households are willing to pay a higher percentage of their income for a favorable climate amenity than the rest of the pollution.[32] Moreover, there exist large geographic variations in estimated values. Retired households distaste the future temperature changes in most climate regions, except for the Northeast region where households are willing to pay for the changes. It implies that, in the majority of urban areas, the benefits from warmer winters are outweighed by hotter summers, while the value of warmer winters exceeds hotter summers in the Northeast region. The positive impact of global warming on future retired households in the Northeast region primarily comes from the mitigation of the coldest climate.

2.6.3 Welfare Evaluation with Mobility and Household Relocations

The previously calculated values for various climate scenarios are conditional on current locations, without considering the potential averting behaviors. Households are assumed to stay in their current MSA and not move in response to temperature changes over time. However, given the heterogeneous preferences for temperature amenities across socio-demographic groups, there can exist further residential sorting driven by the projected climate change over the long

[32]The values of a similar and more friendly climate scenario are found to be around 1%-1.4% of household income for the entire U.S. population (Sinha et al., 2018).

Table 2.7: Temperature Changes and WTP in Current Locations in 2050 and 2100

Region	Midwest	Northeast	West	South	All
Year 2050					
Δ°C in summer temperature	2.9	1.6	3.8	4.2	3.1
Δ°C in winter temperature	2.5	2.9	3.2	3.2	2.9
WTP for the change	-723.6	1,294.9	-1,032.8	-1,516.8	-890.1
Year 2100					
Δ°C in summer temperature	3.1	2.9	4.1	4.6	3.6
Δ°C in winter temperature	2.7	3.6	3.8	3.1	3.3
WTP for the change	-742.8	501.8	-1227.1	-2,112.2	-1.379.4

Notes: The changes in temperatures and WTPs are weighted by the MSA populations in each region. The WTP amounts to the willing to pay for warmer winters, net of hotter summers.

period. Retired households may overcome a moving cost and relocate to another place, even if the adjustments should be relatively rare given the small changes in the temperature. Thus, to calculate an exact welfare measure of temperature changes, I take into account the possibility of migration and allow each retired household to choose the utility-maximizing residence again under new climate scenarios. Given locational attributes and household-specific preferences for temperature-related amenities, the exact welfare change is measured by a household compensating variation (CV), which is implicitly solved in the following equation:

$$U_{ij} = \max_j \left(V_{ij}^0 + \varepsilon_{ij} | Y_i, \mathrm{ST}_j^0, \mathrm{WT}_j^0, \mathrm{VT}_j^0\right) = \max_j \left(V_{ij}^1 + \varepsilon_{ij} | Y_i - CV_i, \mathrm{ST}_j^1, \mathrm{WT}_j^1, \mathrm{VT}_j^1\right), \quad (2.11)$$

where ST_j^0, WT_j^0 and VT_j^0 are current temperature amenities and ST_j^1, WT_j^1 and VT_j^1 are projected future temperature amenities. V_{ij}^1 represents the new utility of household i facing a new climate scenario.[33] CV_i denotes the compensating variation that equals the amount household i is willing to pay in exchange for different temperature-related amenities. Given the location choice and household demographics, the expectation of CV_i becomes:

$$E(CV_i | y_{ij}, Z_{ij}, \mu, \Sigma) = \int CV_i h(\beta_i | y_{ij}, Z_{ij}, \mu, \Sigma) d\beta_i, \quad (2.12)$$

[33]The systematic utility is $V_{ij}^1 = \alpha(Y_i - \widehat{H_{ij}} - Q_{ij} - CV_i) + \mathrm{WT}_j^1 \beta_i^{WT} + \mathrm{ST}_j^1 \beta_i^{ST} + \mathrm{VT}_j^1 \beta_i^{VT} + MC_{ij} + \eta_j$.

where $h(\beta_i|y_{ij}, Z_{ij}, \mu, \Sigma)$ is the conditional probability of preference parameters, β_i. In practice, I randomly draw preference parameters and random part of household utility to compute a number of CV_i for household i in the equation 2.11. Then, I repeat the simulation 100 times and take the average of CV_i to calculate the expected compensating variation across all households.

Table 2.8 reports the expected compensating variations in each scenario across climate regions if all retired households are given full mobility. On average, the average loss of household welfare due to temperature changes is \$602 (1.4% of their annual income) in 2050 and \$993 (2.4% of their annual income) in 2100. The negative numbers equal what a household needs to be compensated for enduring an adverse climate scenario. It can be seen that the preference ranking of various climates remains the same, while the estimates on welfare changes in most cases are lower than WTPs conditional on current locations. The compensation a household requires facing the new climate is on average less than the amount they are willing to pay for staying in the current favorable climate. The differences in estimates arise mainly from the different assumptions on household mobility. When households are free to move, they can improve their welfare by moving to another MSA as long as the utility gain exceeds the generalized moving cost. As a consequence, the actual damage caused by the adverse climate becomes lower due to massive self-adjustments in the residential sorting process.

Table 2.8: Temperature Changes and Compensating Variations in 2050 and 2100

Region	**Midwest**	**Northeast**	**West**	**South**	**All**
Year 2050					
Δ°C in summer temperature	2.9	1.6	3.8	4.2	3.1
Δ°C in winter temperature	2.5	2.9	3.2	3.2	2.9
E(CV) for the change	-413.3	891.2	-738.4	-1,092.2	-602.3
Year 2100					
Δ°C in summer temperature	3.1	2.9	4.1	4.6	3.6
Δ°C in winter temperature	2.7	3.6	3.8	3.1	3.3
E(CV) for the change	-640.3	202.2	-321.3	-1,238.2	-992.5

Notes: The temperature changes and E(CV) are weighted by the MSA population across climate regions. The absolute values of negative numbers are the amounts a household needs to be compensated with a new climate scenario.

Given the full mobility and potential household relocations, the retired population in the U.S. can be geographically redistributed over the years in response to changing climate attributes. Figure 2.6 in Appendix shows the time-variant spatial distributions of the retired population in 2050 and 2100. Specifically, the two maps show the percentage changes of retired households in a local population, compared to the year 2017. It can be seen in Panel (a) that, as a result of taste-based sorting, the Northeast region with relatively cooler summers than average would attract more retirees to reside there in 2050. Many retired households will move north from the South region and southern areas in the West region, driven mainly by the upcoming broiling summers. Some MSAs are expected to accommodate 1.5-2.5% more retired households. These estimates are based on the projected changes in outdoor temperatures. Admittedly, some adaptations, such as installing an air conditioner, can largely mitigate the influences of hotter summers. Due to long-term global warming, the pattern of residential sorting will be further intensified over time. As shown in Panel (b) in 2100, a larger portion of retired households move away from southern California, Texas, and south Florida in the South and West regions, contributing to an overall northbound migration pattern.

Given the high moving cost, rational retired households need to forecast the changing climatic and locational attributes when making a locational decision. If they are not forward-looking and thus cannot fully consider the changing temperature amenities, the model estimation is likely to be biased. Using a dynamic computable general equilibrium model in which households incorporate future stream of utilities with time-variant attributes, some recent papers estimate how the long-run climate changes lead to a population redistribution across the United States (Fan et al., 2018). As a robustness check, I reestimate the household locational choice model with projected future temperatures and current locations. The estimation results do not vary significantly, suggesting that retired households are myopic in some sense. However, the potential biasedness is not as worrying as it seems to be. It is due mainly to the fact that retired households make locational decisions for a shorter time horizon and expect to stay in a residence for at most

20 years during which climate attributes would not substantially change. Therefore, it is still reasonable to estimate the model with recent average temperatures.

2.7 Conclusion

Using the newly released U.S. 2017 census data, this paper documents the relationship between local climate amenities and retired household residential location decisions. The empirical results of the structural sorting model show that climate amenities play an important role in deciding a location in which a retired household chooses to live. It is found that retired households value favorable climate amenities. On average, they are willing to pay $1,209 for a 1°C drop in average summer temperature and $1,114 for a 1°C increase in average winter temperature. The estimated MWTPs by retired households are higher than those by the entire population (Sinha et al., 2018). In addition to the average temperatures, this paper provides the first estimate on the value of another important temperature-related amenity, temperature variability, to account for the particular physical status of retirees. The MWTP for a 1°C decrease in average difference in daily maximum and minimum temperatures equals $486. In the robustness analysis, I find that these empirical results are robust to the specification of economic component but sensitive to the setup of moving cost. In this paper, the economic values for the quality of urban life are reported and found to be higher, compared to the estimates provided by Albouy et al. (2016). Apart from the mean estimates, I use the random coefficient model to investigate preference heterogeneity in climate amenities. The MWTPs for a favorable climate are found to be higher among older retired households with a higher retirement income and disability. Moreover, in terms of geographic variations, this paper confirms that retired households who live in places with a more friendly climate have a higher MWTP for the preferred temperatures more than the average.

In the economics of climate change, many efforts have been made to analyze how projected changes in climate amenities influence daily household activities. To contribute to literature, this

paper also provides some of the first empirical evidence of retired households' WTP to avoid future changes in climate amenities. Conditional on current locations and household preferences for climate amenities, the projected changes in temperatures in the future would cause a welfare loss. On average, households are willing to pay up to nearly 3.3% of their annual retirement income to avoid the upcoming climate scenarios. The calculated household welfare loss caused by unfavorable climate amenities offers much information in the cost-benefit analysis of climate change. Given the large retired population, the large aggregate loss of social welfare motivates the mitigation of global warming.

From an urban planner's perspective, local demographic composition plays a critical role in long-term urban development. The changing climate is important for not only the current generation but also future retired households. It can result in an amenity-driven residential sorting and gradually reshape the geographical distributions of the retired population across localities in the United States. Simulation results forecast the new climate amenities are expected to cause an overall northbound migration of the retired population in the continental United States. This finding is of direct urban policy relevance and has profound implications for local urban planning. Existing literature has shown that retirees are economically beneficial to local areas, due to an increase in the property tax base and a relatively light burden on the public service budget (Duncombe et al., 2001). For policymakers in areas with an upcoming net out-migration, it presents a long-run challenge of providing amenities and other utility enhancing attributes to keep those future retired households. On the contrary, those popular destinations for retirement can start to build more urban facilities, like public parks and nursing homes, to accommodate more retired households.

2.8 Acknowledgement

Chapter 2, in full, has been recently accepted for publication. Jiajun Lu. The dissertation author was the sole author of this chapter.

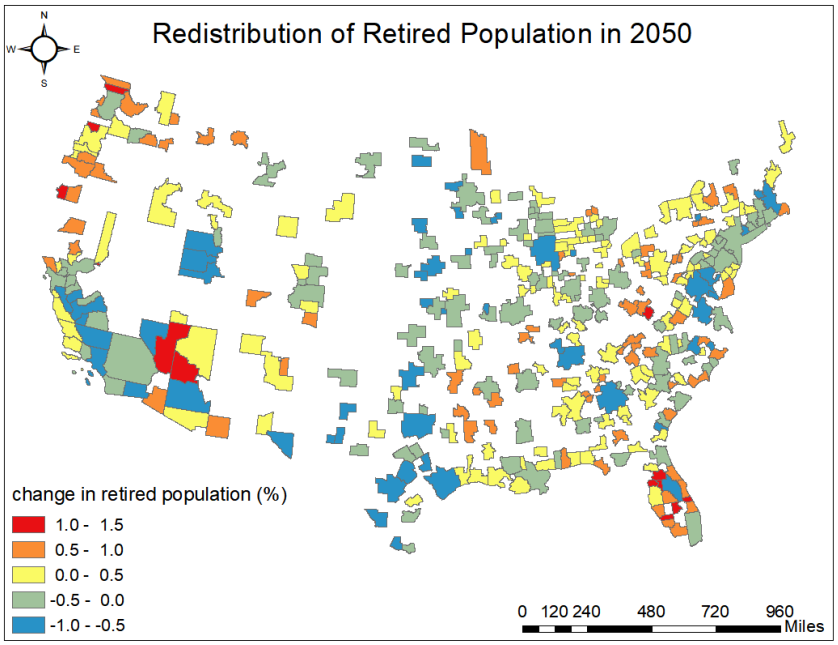
2.9 Appendix

2.9.1 Supplemental Table and Graph

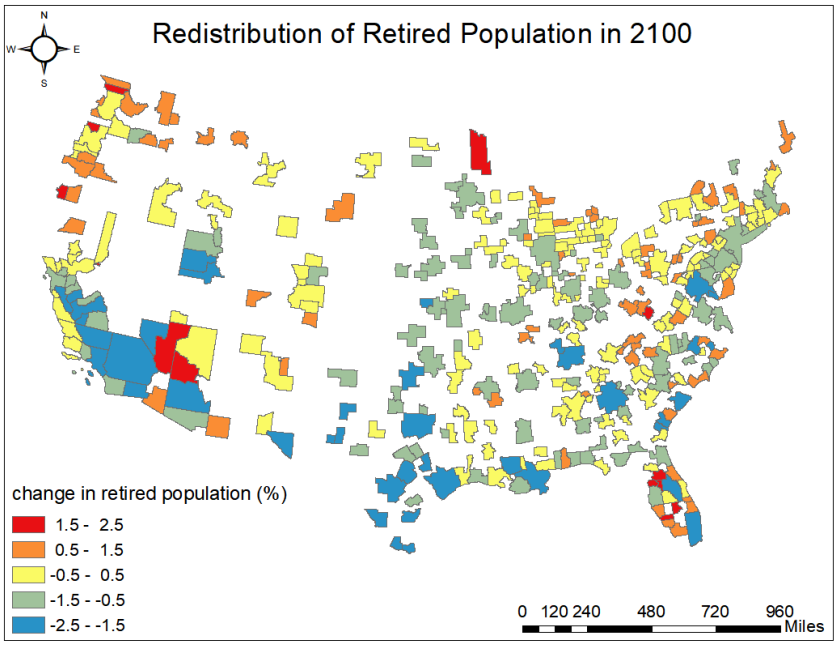
Table 2.9: Summary Statistics of the Coefficients on Hedonic Housing Equations

Dependent variable: $\ln H_{ij}$, the regression equations (2.2) for 377 MSAs					
Variable	**Mean**	**Std Dev**	**Variable**	**Mean**	**Std Dev**
Owner-occupied unit	0.038	0.012	Single-family house	0.129	0.144
Complete kitchen	0.109	0.121	Complete pluming	0.025	0.032
Running water	0.082	0.125	Bathtub or shower	0.077	0.178
1-bedroom unit is left out as reference					
2 bedrooms	0.157	0.139	3 bedrooms	0.233	0.127
4 bedrooms	0.339	0.287	$\geq$5 bedrooms	0.539	0.163
House with less than 1 acre is left out as reference					
House with 1-10 acres	0.131	0.268	House with $\geq$10 acres	0.277	0.267
Detached and attach one-family house is left out as reference					
2-9 units in structure	-0.093	0.276	10-19 units in structure	-0.177	0.259
$\geq$20 units in structure	-0.233	0.237			
Unit built in 2017 is left out as reference					
2-5 years old	-0.043	0.345	6-20 years old	-0.153	0.274
$\geq$20 years old	-0.098	0.298			
Total N = 1,203,865, average adjusted R^2 = 0.1519					

Notes: This table displays the summary statistics of hedonic housing coefficients from the equation (2.2) for all MSAs with a total of 1,203,865 housing units. The Mean and Std Dev denote the mean and standard deviation of coefficients from 377 MSAs.



(a) Geographical redistribution in 2050



(b) Geographical redistribution in 2100

Figure 2.6: Geographical Redistributions of Retired Population in 2050 and 2100

Table 2.10: Sensitivity of MWTP for Temperature Amenities to Alternative Model Specifications

	Model 1 Base Model		**Model 2** $\alpha(Y_i - \widehat{H_{ij}}) + \beta Q_{ij}$		**Model 3** Alternative MC_{ij}	
Variables	Estimate	MWTP	Estimate	MWTP	Estimate	MWTP
Hicksian bundle, α	4.7621e-5		4.0987e-5		3.2143e-5	
Hicksian bundle, β			-4.2113e-5			
Mean summer temperature	-0.0576	-1,209.97	-0.0527	-1,287.09	-0.0261	-809.80
Mean winter temperature	0.0531	1,114.09	0.0449	1,097.11	0.0253	789.12
Variability in temperature	-0.0231	-486.13	-0.0175	-427.18	-0.0124	-388.57
Moving out of metro	-0.0801	-1,683.18	-0.0643	-1,569.98	-0.0102	-317.31
Moving out of state	-0.0426	-886.76	-0.0449	-1,097.03	-0.0093	-287.93
Moving out of region	-0.0198	-416.09	-0.0245	-597.90	-0.0022	-67.09
Moving distance	-3.8706e-4	7.9811	0.0002	6.0129		
Moving distance squared	-5.2321e-5	1.0987	3.2347e-05	0.7892		

Notes: The MWTPs are measured in 2017 U.S. dollars. The Model 2 relaxes the linearity in the Hicksian bundle but maintains the setup of moving cost. In Model 3, the alternative moving cost becomes $MC_{ij} = \pi_1 I_{ij}^{\text{Metro}} + \pi_2 I_{ij}^{\text{State}} + \pi_3 I_{ij}^{\text{Region}}$, keeping the linearity in the Hicksian bundle.

Chapter 3

Selling the Modern Residence: A Tale of Zestimates and Open Houses

3.1 Introduction

Given the massive size of the U.S. residential real estate market, which is only surpassed by the overall stock market, there is surprisingly little work by economists on how that market works at the ground level. This stands in contrast to the huge amount of economic work focusing on the implications of the market collapse that triggered the Global Financial Crisis (GFC) and a whole host of interesting microeconomic issues where houses are playing a leading or major supporting role. The underlying process of selling a home, since then, has changed radically over the last decade. For roughly a hundred years, most homes in the United States were bought and sold using real estate agents who were members of multiple listing services. The listing service simultaneously maintained a master list of residential properties for sale, including the listing agent, price, and other characteristics such as the year the structure was built, the number of bedrooms, lot size, and the school district. It also served as the storehouse of information on past home listing and sales used by a seller's agent to help set the asking price for the home and by a

buyer's agent to assess the competitiveness of that price. The advent of online real estate websites and, particularly, Zillow, have fundamentally changed the structure of how the U.S. home sales are made. We characterize how the process of selling and buying homes now works.

The vast majority of homes are still bought and sold using real estate agents who split a percentage of the sales price, and over 600 multiple listing services (MLS) serving specific areas exist. What has changed is that, after years of different degrees of conflict and cooperation between individual MLS and online websites, MLS now supplies those websites with a list of homes that are on the market for sale. While the details of how this came to be differed by location, one of the contributing factors was undoubtfully the consolidation of the major online websites. The Zillow Group, which includes Zillow, its former major competitor Trulia as well as other smaller sites such as Hotpad, and Realtor.com now control most of the U.S. internet traffic.[1] Most home sellers and buyers now search online for information related to undertaking the transaction from finding agents to locating times at which homes for sale were open for public viewing.[2] What is clear is that the one-time close to a monopoly that local MLS had on information about what homes were for sale is no longer the case.

These online websites have also broken the near-monopoly that local real estate agents and their MLS once had on the information needed to make informed decisions about any particular home value. Most prominently, Zillow provides its "Zestimate" of the market value of almost every home in the United States, which is based on a proprietary hedonic pricing model.[3] For many participants in the housing market, looking at how the Zestimates of your house and your neighbors' houses change over time has become a national pastime. Our attention will be focused on the role that this independent estimate of home value plays in how selling realtors, who still possess superior knowledge of a home's physical condition and seller characteristics such as

[1] Realtor.com is an online platform on which the real estate agents themselves launched their trade group, the National Association of Realtors.

[2] A 2018 study by the National Association of Realtor, for instance, finds that 82% of home buyers seek information on what homes are available for sale on the internet. https://www.nar.realtor/sites/files/documents/2018-real-estate-in-a-digital-world-12-12-2018.pdf

[3] https://www.zillow.com/Zestimate/

degree of impatience, set a list price. We also investigate how factors that predict a divergence between the Zestimate and the home's list price and how the divergence between these two indicators influences a home's eventual sales price and the number of days it is on the market.

There are a whole host of other decisions that home sellers need to make. One is the choice of selling agent. Zillow provides information on the agent's number of years of experience and a numeric agent rating, which is mainly based on surveys of past home buyers and sellers who have had interactions with individual agents.[4] We look at how these publicly available indicators predict various aspects of the home sale.

Another is the decision on whether to hold an initial open house where the time and date when the home is available for public viewing is preannounced. Open houses, which are expensive for a realtor to hold, allow attendees to obtain a more complete assessment of a home's condition and surroundings than is available from online sites and agent fliers. While the holding of open houses has long been a tool used by real estate agents, the nature of this tool has substantively changed with the advent of Zillow and Realtor.com. Once a haphazard affair driven by realtor flags and signs on key street corners in a neighborhood as well as occasional advertising in newspapers and real estate magazines, potential homebuyers can now obtain online a relatively complete listing of home houses with the ability to define spatial areas by neighborhood or zip code and to filter results by housing characteristics such as the number of bedrooms and price. We look at the factors that predict whether a seller's agent hosts an open house when the home first goes on the market, which tends to attract more in-person viewers of the home in the short window when it is considered "new" on the market and the role that decision plays on the major home sales statistics. Because open house dates are chosen before weather realizations, we can also look at how those random realizations, which economists have examined in several other contexts, influence a home's eventual transaction price and time on the market.

Our empirical analysis is conducted with over 220,000 single-family home sales from

[4] https://premieragent.zillow.com/re/agent-reviews-and-ratings-faq/

Zillow. The sample consists of all single-family homes that first appeared in Zillow for sale in 04/2018-08/2018 across 314 U.S. cities with a population of a hundred thousand or larger. We then follow these houses for two years to determine their sales prices if they are sold. Information on homes going on the market in these cities was scrapped on a daily basis, which allowed us to obtain open house information and initial asking price.

We find that the list price of a house is typically closer to its Zestimate than its sales price. Moreover, if the list price relative to its estimated value is set to be lower, the property is more likely to be sold at an over-the-list price. Hosting an open house when the home first goes on the market typically brings a price premium, increases the probability of selling a property above the initial list price, brings a price premium relative to the home's Zestimate, and reduces its selling time, holding other factors equal. The positive impact of an open house is larger for dwellings sold in the warmer season of the housing market. Adverse weather realizations, such as rain on an open-house date, tend to reduce the sales price and increase time on the market. Overall, a rich picture emerges from our analysis of how single-family homes in the U.S. are bought and sold at the individual level.

The remainder of this paper is organized as follows. Section 2 presents a brief review of relevant literature. The Zestimate and key tools available to a seller's realtor, including the list price for a home and whether to hold an initial open house, are introduced in section 3. The data used in our empirical work and our methodology framework are described in sections 4 and 5, respectively. Empirical results are presented in section 6. The paper concludes with a summary of key findings.

3.2 Literature Review

Researchers have long questioned whether or not hosting an open house impacts the outcomes in residential real estate markets, but only a few have empirically examined its effects.

Prior research mainly focused on the use of an open house and its influence on sales performance, e.g., time on the market and ultimate selling price. Firstly, the effects of an open house on time-on-the-market and probability of sale are ambiguous. Rutherford et al. (2005) analyze how a public open house works as a potential determinant of a dwelling's selling time and find that a home's marketing time can be reduced by the use of open houses. On the contrary, Huang and Rutherford (2007) also provide similar empirical evidence of selling time, showing that having an open house increases the marketing time. As to sales price, the effect of an open house is less uncertain. In the classic auction theory, a dwelling tends to be sold at a higher price given more potential bidders if holding an open house (Milgrom and Weber, 1982). Meanwhile, relevant empirical evidence has been provided, suggesting that holding an open house brings a price premium and this marketing strategy has a positive price effect (Allen et al., 2015).

Apart from an open house, another important marketing strategy, setting a list price, has a significant influence on sales performance. Much research has been done to analyze the listing price strategy in the real estate market. Horowitz (1992) completes one of the first studies and provides both theoretical and empirical analyses of the impacts of a listing price. It explains why list prices are price ceilings that preclude the possibility of sales at higher prices. McGreal et al. (2009) examine different factors that influence the price premium and find that the majority of sales of residential properties are at a premium to the list price in the United Kingdom. Han and Strange (2016) also develop a model showing that the asking price plays a directing role in buyer search behavior and is of great relevance to transaction prices.

On top of the two commonly used marketing strategies mentioned above, real estate agents in the housing market start paying attention to home value estimates in recent years. Since the release of Zestimate by Zillow, several recent papers in economics have sought to explain how it would fundamentally influence the residential housing markets (Rascoff and Humphries, 2015). As opposed to a large body of literature on open house and listing price, there are only a few studies on Zestimates, of which the major focus is to investigate whether the estimates are

accurate and can make real estate more transparent. Hagerty (2007) first evaluates the accuracy of Zestimates and believes that Zillow's Zestimates often are very accurate. In contrast, Corcoran and Liu (2014) argue that these home values estimated by Zillow are often overestimated. Gan et al. (2015) introduce the Zestimate in detail and examine the machine learning techniques to attain Zestimates. Swango (2015) also summarizes the main database of Zestimates and suggests that these references are very useful to real estate analysts and appraisers. Frey et al. (2013) adopt Zestimates to predict the value of urban wetland and results show that the estimated values are predictive of the eventual sales price in a hedonic pricing analysis. Given its large influence on the modern housing markets, the potential impact of Zestimate on list price strategy and real estate sales performance is an issue worthy of additional research.

3.3 Empirical Background

3.3.1 Role of Zestimate

Before the availability of Zestimate, the U.S. residential real estate has long been a market where the information is, to a large degree, privately held, making home buyers and sellers much less informed than their agents. Normally, a selling agent sets a list price with superior knowledge of home characteristics and the actual property value of a dwelling, while a buying agent only has the information about what houses in the neighborhood are being sold. However, due to the publicly available Zestimate taken as the benchmark, the list price set by the seller's realtor implies whether the house has more or less value than its Zestimate.

Since the Zestimate has been getting more attention in the housing market nowadays, the asymmetric information in the real estate market gradually reduces. Therefore, this high-quality public signal radically changes the U.S. residential housing market, which is later introduced in detail in the empirical models.

3.3.2 Open House and List Price

To promote real properties, real estate agents usually implement some active marketing strategies. One such strategy is to schedule a public open house event during which the general public can enter the property and examine its characteristics on site. A broker will advertise a public open house in Zillow, in addition to other real estate websites. Another important approach to marketing a house is to set an initial list price that helps either raise sales price or expedites the selling process. In this paper, we mainly focus on the two typical marketing strategies, i.e., holding an open house and setting an optimal list price.

Figure 3.1 illustrates the timeline for selling a dwelling on the residential real estate market, where its Zestimate is available to the public, and the strategies of both an open house and list price are probably implemented. Typically, before listing a housing unit, a seller's real estate agent observes its detailed characteristics, Zestimate, and community attributes. Given the information on the dwelling, the agent sets the initial list price and decides whether to hold a speedy open house in the first week at the same time. Then, potential buyers bid for the house at the early stage, during which a deal is likely to be made, and the sales price is observed. If there is no deal, the house remains unsold on the market.

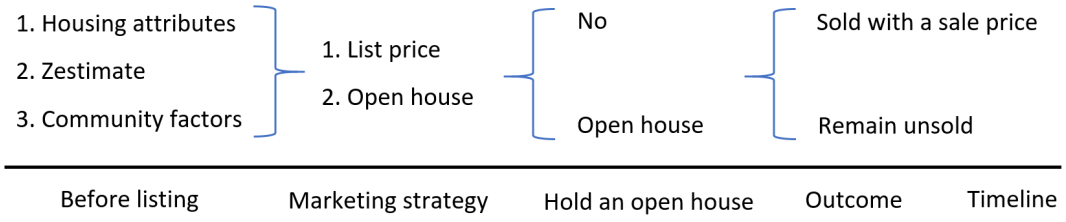


Figure 3.1: The timeline for selling a house on the market

3.4 Data

3.4.1 Data Sources

This paper is conducted with a unique and comprehensive dataset, comprised of housing transaction data, dwellings and agent characteristics for houses both sold and unsold on the market, and critical climate and location attributes that influence sales performance.

Housing Transaction Data

The main source of data for our analysis is Zillow, one of the primary databases for residential real estate in the United States.[5] It provides much detailed information in the selling process and has become an online real estate marketplace that seeks to make information available about all dwellings.

The relevant variables are divided into two categories, i.e., prices and timing variables. The first involves all critical price information, including eventual sales prices and initial list prices.[6] In addition to the two typical prices, Zillow provides another price, Zestimate, an estimated market value of a residential property. It incorporates public data from government records, e.g., tax assessor, on assessed value, home characteristics, past sales, neighborhood characteristics, and market conditions. In many instances, it involves more detailed information from previously initiated realtor sales efforts for either the current or previous homeowner and owner-initiated changes to home characteristics, which often reflects remodeling efforts or correction of errors from other information sources. A home's Zestimate is based on a proprietary algorithm that can be viewed as a machine learning version of a sophisticated hedonic pricing model (Gan et al., 2015). It has become more accurate over time, as more information on prediction errors on each property is amassed and differences between a property's sales price and its Zestimate are

[5]Zillow has data on approximately 110 million homes across the United States (Corcoran and Liu, 2014).

[6]Admittedly, in some non-disclosure states, like the State of Texas, Zillow has no legal right to release the data to the public. However, it obtains information on actual sales prices from a number of sources, including, but not limited to, the Houston Association of REALTORS (HAR). See: https://www.har.com/

continually observed (Corcoran and Liu, 2014). The second category of housing data relates to critical timing variables during the whole selling process, including the date a dwelling unit is initially listed and the date when it is sold if applicable. Then, based on the two dates, the length of time the house is put on the market is calculated.

Dwelling and Agent Characteristics

As the largest online real estate database company, Zillow provides detailed information on home attributes, including numbers of bedrooms and bathrooms, floor area, lot size, storeys in a dwelling, the number of parking spots, year built, and a cooling system. In the transaction records, we also attain the information of a home's exact address and characteristics of the seller's agent, including their ratings and work experiences. A realtor's rate depends on how active an agent is on the market. All real properties in the sample are identified by a unique identifier in Zillow.

Open House and Weather Conditions

In addition to housing and agent characteristics, this paper collects another timing variable of holding an open house, the marketing strategy that also influences sales performance in the selling process. Zillow records accurate information of the time when an open house is held, including both hours and dates.

We select a speedy open house held in the first ten days after a house is listed since it is the minimum period that covers at least a whole weekend when an open house is usually held. A speedy open house is also restricted to this time period to get rid of endogenous open houses conditional on past experience of sales performance on the market. It is due to the fact that an open house that happens a long time period after being listed is likely to be influenced by an observed sales potential on the market. The agent may experience a long-term wait with very few potential buyers offering an acceptable bid price and decide to schedule an open house later.

On the contrary, if a real property agent hosts an open house a short period after being listed, the decision is most likely prearranged, resulting in an exogenous impact on sales prices.

Given the accurate timing of an open house, we extract associated short-run weather conditions to match the period of an open house. The weather information comes from National Oceanic and Atmospheric Administration (NOAA) of the United States, including daily mean temperatures, wind speed, precipitation rate, snowfall, and sunshine.[7] These short-term exogenous weather conditions have been examined and found to impact the sales performance in other studies (Simonsohn, 2010; Busse et al., 2015). Thus, this paper also analyzes the potential influence of weather conditions in open house days on the sales performance of dwellings. Specifically, due to the potential large variations in weather condition within a day, this paper adopts the hourly average of weather data, rather than the daily average. The timing variables recorded on an hourly basis substantially improves the accuracy of the model estimation related to an open house.

3.4.2 Study Area and Sample Selection

To make our analysis broadly representative of the U.S. home sales, we take the entire U.S. real estate market as the study area. Due to data limitations, we place some restrictions in the sample and clean the data for practical purposes.

First, we confine our analysis to geographic areas where the residential real estate market should be thick and well functional in the sense of many agents, many buyers, and many sellers by looking at single-family homes. This is done by using all transactions posted in Zillow in 314 cities that U.S. Census Bureau reports as having a population of a hundred thousand or more.[8] Figure 3.2 illustrates the map of the cities in the study area, covering more than 62.6% of the U.S. single-family homes.[9] A substantial fraction, nearly 18.2%, of the remaining single-family homes are in smaller cities that are part of nearby metropolitan areas. Hence, we would expect

[7] https://www.ncdc.noaa.gov/cdo-web/datasets

[8] https://factfinder.census.gov/faces/tableservices/

[9] https://www.census.gov/hhes/www/housing/census/historic/units.html

sales there to behave in a similar manner.[10] Secondly, we look only at single-family home sales. This excludes condo apartments in multi-unit structures but still leaves a large number of condo townhouses, duplexes as well as free-standing home.[11]

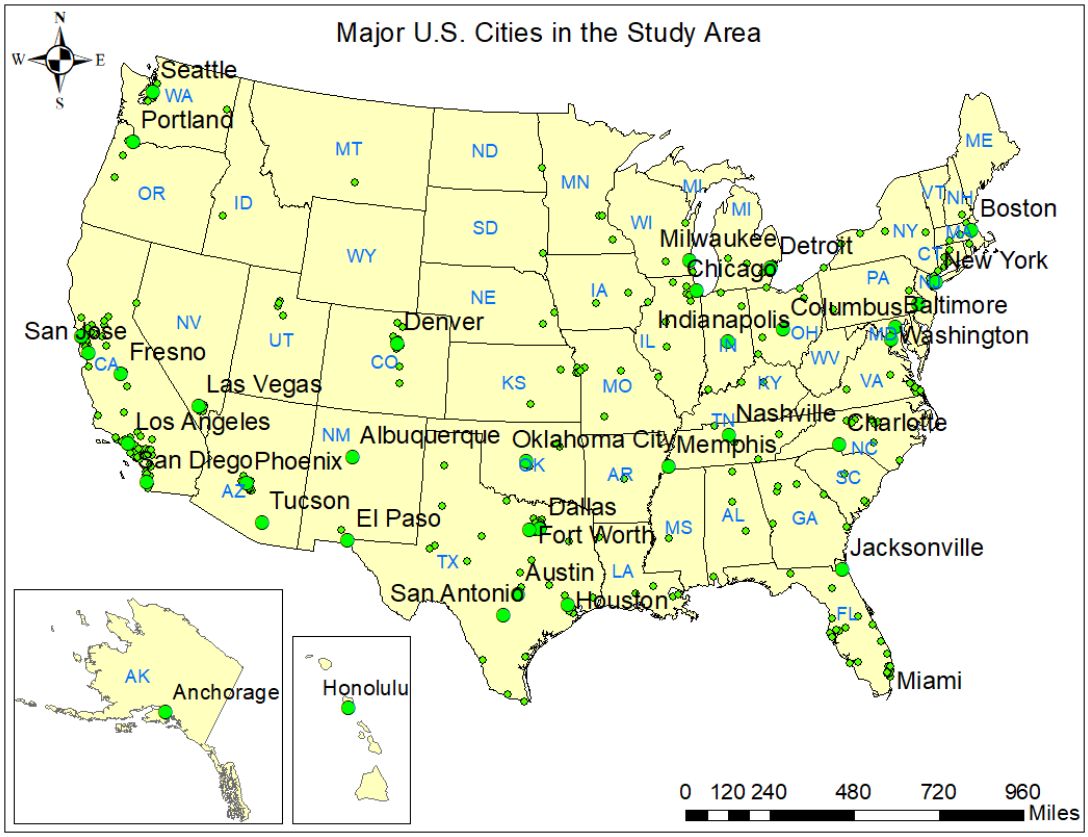


Figure 3.2: Map of the U.S. cities in the study area. It displays the 314 cities that U.S. Census Bureau reports as having a population of a hundred thousand or more in 2018. The size of green dots is proportional to the number of single-family houses in each city.

To produce reliable summary statistics, we clean the sample data in multiple ways. Incomplete transaction records with missing information of critical attributes are first removed.[12] Then, we drop the data of houses that are either too cheap or expensive. List prices are taken

[10]The recent work of Piazzesi et al. (2020) showing that residential real estate search patterns for smaller cities in the San Francisco Bay area are different from the area's central city suggests that it would be useful to specifically look at smaller cities and towns within metropolitan areas to see how well our results carried over.

[11]Condo apartment units contained in large buildings often face a different real estate selling process and involve additional attributes that do not influence single-family home sales.

[12]The observations flagged as gross outliers tend to involve houses with substantial missing information on key variables like square footage or involve situations where the specific property for sale at an address is unclear.

as the reference for possible outliers, due to the fact that list prices are much less likely to be apparently misreported. A seller's realtor has more inside information on a property's value and thus less likely sets the list price wildly wrong, even if there is a substantial problem with Zillow (Corcoran and Liu, 2014). The houses whose prices are set too low might have some hidden defective housing attributes that bias the Zestimate and sales performance. Cheap houses that are in need of extended renovations can more likely be sold to real estate agents who are about to put extra efforts into the renovations. As for the high-priced housing units, they can also be sold to unusual real estate agents who consider it as an investment. For instance, they could tear down the big mansion and construct multiple single-family houses on the same land. Therefore, to eliminate the influential outliers, we trim off the lowest and highest 1% of list prices in the data to conduct the main empirical analyses in this paper, which has been widely adopted by other studies (Temple, 2000; De Haan, 2007).[13]

Finally, some further data cleaning work is done based on an observed housing attribute, the heating system. The presence of a heater is required in some states. If it does not exist, the dwelling unit is more likely to be sold to real estate agents who provide renovation and remodeling, rather than regular homebuyers. In another case, the house could be too old or in relatively bad conditions. It is found that there are nearly 1.2% of dwellings without a heater in the data. These uncommon houses without a heating system are then dropped out of the sample in the estimation.

3.4.3 Summary Statistics

Given the sample selection and data cleaning, we obtain the sample data with a total of 220,218 single-family houses over the period 04/2018-03/2020, among which 218,147 dwellings have been sold, and the other 2,071 houses are still unsold as of 03/2020. All houses are listed

[13]To check the robustness of model estimations without outliers, we also estimate the models with the exclusion of 2% and 3% of observations on each side of the distribution and find that the main results are not largely altered.

over the period 04/2018-08/2018, and, after tracking all houses in the following two years, nearly 99% of them have been sold.

Table 3.1 presents the summary statistics for all variables used in the empirical study. In the sales information, initial list prices are the first asking price of real property, while Zestimates are recorded the same day on which real property is listed. It is seen that sellers' real estate agents set an initial list price slightly higher than its Zestimate, while Zillow, on average, overestimates the home values. Figure 3.3 illustrates the relationships among list prices, Zestimates, and sales prices. Panel (a) presents the response surface plot, showing that sales prices consistently increase with Zestimates, while the list price relative to Zestimate influences an eventual transaction price with a less clear pattern. Panels (b-d) display the binary relations of the three prices. We estimate a weighted linear least-squares regression using the first-degree polynomial and draw the locally weighted scatterplot smoothing (LOWESS). Panel (b) shows that most dwellings are sold at a price lower than its initial list price. As presented in Panel (c), a majority of houses are overestimated by Zillow. Nearly 69% of dwellings have a sales price lower than its list price, while over 84% of houses have a Zestimate higher than its eventual transaction price. In Panel (d), it is observed that list prices are much closer to Zestimates than actual sale prices, suggesting that realtors refer to Zestimate when setting an initial list price. Figure 3.6 in Appendix also illustrates the bivariate kernel regression plots of the sales price on Zestimate and sales price on the list price, respectively. The kernel density estimations are implemented with the Nadaraya-Watson algorithm using a Gaussian kernel and default optimal bandwidth. It is shown that list prices are less spread out and closer to sales prices than Zestimates, which confirms that a seller's agent typically obtains some unobserved information about the house to be sold even if Zestimates capture most price information.

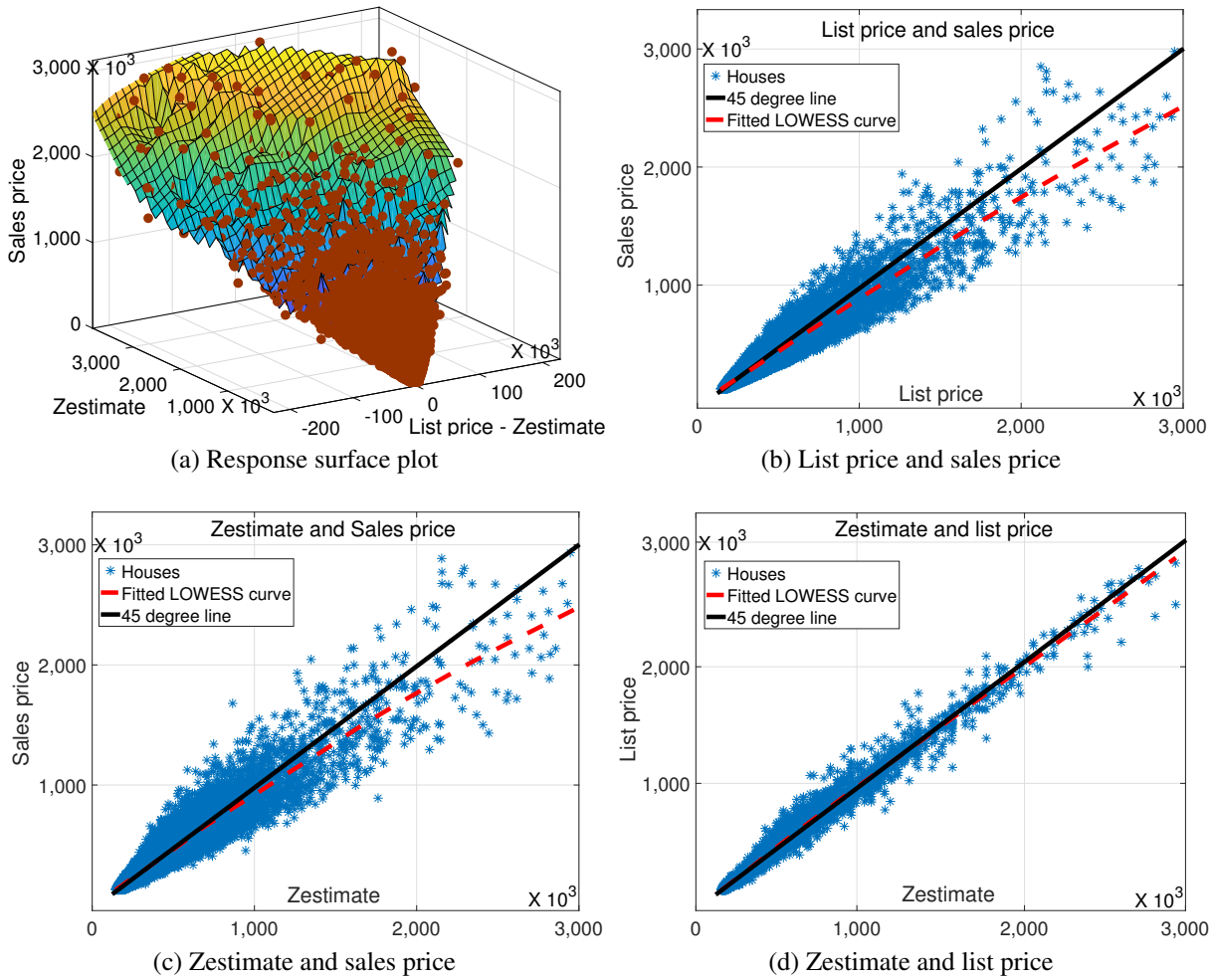
Apart from the price information, we find that the average selling time of these houses is 88.5 days in the residential housing market.[14] It is seen that around 11% of agents host an

[14]The percent of houses being sold after each number of days and cumulative survival probability on the market are presented in Figure 3.5 below.

Table 3.1: Summary statistics of the variables

Variable	Description	N	Mean	SD	Min	Max	Source
Sales information							
salep	Sales price	218,147	457,365	407,594	106,879	2,846,814	Zillow
listp	Initial listing price	220,218	472,740	428,509	108,377	2,951,725	Zillow
Zestimate	Value estimated by Zillow	220,218	464,592	470,466	107,644	2,863,066	Zillow
time	Days on the market	220,218	88.54	53.64	4	647	Zillow
open	Open house in first 10 days	220,218	0.1135	0.3172	0	1	Zillow
sold	If the house is sold	220,218	0.9906	0.0965	0	1	Zillow
Dwelling attributes							
bedroom	Number of bedrooms	220,218	3.4097	0.8501	1	6	Zillow
bathroom	Number of bathrooms	220,218	2.4017	0.8472	1	4	Zillow
floor	Floor area (square feet)	220,218	1,990.12	1,113.26	500	9,000	Zillow
lot	Lot size (square feet)	220,218	9,121.07	6,130.16	500	29,400	Zillow
yard	Yard size (square feet)	220,218	6,387.12	2331.63	0	20,400	Zillow
storey	Storeys in a dwelling	220,218	1.4013	0.8172	1	3	Zillow
age	Age of dwelling	220,218	38.74	25.19	1	78	Zillow
cooling	Access to a cooling system	220,218	0.8378	0.3686	0	1	Zillow
Agent characteristics							
rate	Rating of seller's realtor	220,218	4.6	1.2	1.0	5.0	Zillow
exp	Work experience (years)	220,218	6.4	3.7	1	16	Zillow
Weather conditions on periods of an open house							
temp	Temperature (°C)	24,994	17.32	5.21	-10.41	37.35	NOAA
prep	Precipitation (inch)	24,994	34.12	11.42	9.55	64.09	NOAA
snow	Snowfall (inch)	24,994	20.9	21.44	0.00	89.13	NOAA
wind	Wind speed (miles/hour)	24,994	9.75	2.84	6.23	12.53	NOAA
sun	Percent of sunshine (%)	24,994	62.44	7.35	39.01	85.03	NOAA
unem	unemployment rate (%)[1]	7,536	4.4026	3.4351	2.1	15.8	BLS

Note: The summary statistics of variables are calculated over 04/2018-03/2020 on a daily basis, except for the weather data calculated on an hourly basis on each date. The economic variables are measured in the 2020 U.S. dollar. [1] The unemployment rates are monthly estimates for all 314 cities over the entire study period.



(a) Response surface plot

(b) List price and sales price

(c) Zestimate and sales price

(d) Zestimate and list price

Figure 3.3: The graphs illustrate the relationships among list price, Zestimate, and sales price. Prices are measured in 2020 U.S. dollars. We estimate a weighted linear least-squares regression using a first degree polynomial and draw the locally weighted scatterplot smoothing (LOWESS).

open house in the first ten days. Table 3.1 also shows the summary statistics of the dwelling and agent characteristics. We calculate yard sizes of the houses according to the information of floor area and lot size.[15] Moreover, there are a total of 24,994 open houses held in the sample, and the corresponding short-run weather conditions are measured during the open houses. As an important location attribute that influences how fast a house can be sold on the market, the data of local unemployment rates are also included in the dataset (McGreal et al., 2009). The monthly estimates of city-level unemployment rates for all 314 cities over the 24 months in the study period are provided by Google Public Data Explorer that collects from the U.S. Bureau of Labor Statistics.[16]

3.5 Methodology

This section introduces the methodology and empirical strategies applied to examine how houses are sold via Zillow in the modern U.S. residential housing market.

3.5.1 Accuracy of Zestimates and List Price Signaling

To explore the accuracy of Zestimates, we estimate several models to decompose variations of sales prices into a couple of factors. We first explore how much Zestimates, per se, are predictive of sales prices by estimating the baseline model that regresses sales price on Zestimate at level terms without intercept as follows:

$$\text{salep}_{ic} = \text{Zestimate}_i \beta_1 + \varepsilon_{ic}, \tag{3.2}$$

[15] The yard size of a house is calculated as follows:

$$\text{Yard size} = \text{Lot size} - \frac{\text{Floor area}}{\text{Number of storeys}}, \tag{3.1}$$

where the areas are measured in square feet.

[16] https://www.google.com/publicdata/directory

where the dependent variable, salep_{ic}, is the sales price for dwelling unit i in city c and the single regressor, Zestimate_i, denotes the value of dwelling i estimated by Zillow on the listing date. ε_{ic} is the idiosyncratic error term to be clustered by city.

Apart from Zestimates, we further quantify how much extra information realtors bring by incorporating an initial list price relative to its Zestimate. The following model is estimated to explore the effect of list price signaling on the sales price:

$$\text{salep}_{ic} = \text{Zestimate}_i\beta_1 + (\text{listp} - \text{Zestimate})_i\beta_2 + \varepsilon_{ic} \tag{3.3}$$

where $(\text{listp} - \text{Zestimate})_i$ denotes the difference between the list price and its Zestimate, where listp_i is the initial listing price of the dwelling i.[17] After estimating the baseline model, we are able to explore how much predictive power Zestimates have for the reasonable market values and how the difference between a list price and Zestimate impacts the eventual transaction prices.

According to the specification of the model 3.3, the signaling effects are restricted to be symmetric between above and below the Zestimate. We relax the symmetric restriction and allow the asymmetric signaling effects of a list price on its sales price. The open house and agent characteristics are incorporated to provide additional information in explaining the variations in sales prices. The model with an open house, agent characteristics, and asymmetric signaling effects of a list price is estimated without intercept as follows:

$$\begin{aligned} \text{salep}_{ic} =& \text{Zestimate}_i\beta_1 + \mathbb{1}(\text{above})_i \times (\text{listp} - \text{Zestimate})_i\beta_2 + \\ & \mathbb{1}(\text{below})_i \times (\text{Zestimate} - \text{listp})_i\beta_3 + \mathbb{1}(\text{open})_i\beta_4 + \quad , \\ & \text{rate}_i\beta_5 + \text{exp}_i\beta_6 + \varepsilon_{ic} \end{aligned} \tag{3.4}$$

where the indicator variable, $\mathbb{1}(\text{above})_i$, equals 1 if the listing price is set greater than or equal to its Zestimate, while $\mathbb{1}(\text{below})_i$ equals 1 if the list price is lower than its Zestimate. β_2 and

[17] An initial list price is chosen for the listing price strategy, for that it cannot be influenced by past experience.

β_3 represent the different signaling effects of a list price in two opposite ways. The indicator variable, $\mathbb{1}(\text{open})_i$, equals 1 if an open house is offered in the first 10 days on the market. rate_i and exp_i represent ratings and work experience in years for the realtors, respectively.

3.5.2 Probability of Sale

As two key indicators of sales performance, the probability of sale and selling time are influenced by marketing strategies. We first aim to examine what determine if a house can be sold. Since unobservable factors that vary by locations affect both the decision to hold an open house and how fast houses can be sold, such as social norms among local realtors, the selection of an open house is likely to be endogenous. Therefore, we need to examine what determines the choice of an open house made by the seller's agent and address the endogeneity in a decision to hold an open house, when exploring the impact of holding an open house. To this end, we adopt the proportion of holding an open house in a city as an instrumental variable (IV) in estimating binary choice models. The instrumental variable is reasonably exogenous since the percent of holding an open house in an area is preexistent and determined by the local social norm, rather than other unobserved locational attributes that possibly influence the probability of sale. Specifically, a two-stage Probit model with the endogenous regressor is estimated using maximum likelihood estimation and robust standard errors clustered at the city level as follows:

$$\begin{aligned} \mathbb{1}(\text{sold})_{ic} &= \begin{cases} 0, \ y^*_{ic} < 0 \\ 1, \ y^*_{ic} \geq 0 \end{cases} \\ y^*_{ic} &= X_{ic}\Gamma + \mathbb{1}(\widehat{\text{open}})_{ic}\beta_1 + \text{unem}_c\beta_2 + \mu_{ic} \\ \mathbb{1}(\text{open})_{ic} &= X_{ic}\Pi + \text{opencity}_c\lambda + \nu_{ic} \end{aligned}, \tag{3.5}$$

where $\mathbb{1}(\text{sold})_{ic}$ is an indicator variable that equals 1 if the house i in city c is sold on the market. X_{ic} is the vector of exogenous house-specific variables, including prices, dwelling attributes, and

agent characteristics, while unem_c is the unemployment rate in the local city c. The endogenous variable, $\mathbb{1}(\text{open})_{ic}$, is instrumented in the first stage by exogenous regressors, X_{ic}, and opencity_c, the proportion of holding an open house in a city. By assumption, ν_{ic} and μ_{ic} are the error terms in the two regression models and follow the joint normal distribution $(\nu_{ic}, \mu_{ic}) \sim N(0,\ \Sigma)$, where σ_{11} is normalized to 1 to identify the model.

In the vector X_{ic}, we select all defined variables and calculate some other key factors relevant to the probability of sale as follows. A list price that is set much higher than the median sales price could reduce the likelihood of being sold. We first obtain the median sales price in a local city and compute the percentage difference between the list price and the city-level median sales price, i.e. $\text{listoversalesp} = \frac{\text{listp}-\text{medsalep}}{\text{medsalep}}$, where medsalep denotes the median sales price in the city. Similarly, if the list price is set higher than its Zestimate, it is supposed to be less likely for the house to be sold. The percentage difference between the list price and its Zestimate is calculated as $\text{listoverZest} = \frac{\text{listp}-\text{Zestimate}}{\text{Zestimate}}$. To control for nonlinearity in the influence of prices, the quadratic forms of the two variables computed above are included as well. We then explore the influence of the number of bedrooms, due to the fact that, if a house has many bedrooms, it would be the target house by fewer potential buyers in the local area. So, we look at the numbers of bedrooms and attain the 90th percentile for the number of bedrooms in a local city. $\mathbb{1}(\text{manybed})_{ic}$ equals 1 if the house i's number of bedrooms is in the 90th percentile in the local city c. The vector also involves other covariates, including the cooling system, if a house has more than one storey, yard size, rating, and work experience of the seller's agent.

3.5.3 Survival Analysis

In addition to the probability of sale, survival analysis is conducted to investigate what influences the length of time periods during which the houses are on the housing market using survival analysis methods. According to the real estate data, there are still some houses unsold as of the end of the study period, which truncates the data of selling time for these observations.

Therefore, to control for the subsample of unsold houses, the survival analysis is performed with censored survival data. Specifically, we adopt a proportional hazard regression model, also named Cox regression, with the key assumption that the probability for any house to be sold is a fixed proportion of the hazard for the baseline house at any time. It means that the hazard ratio depends only on the predictor variables and not on time (Miller Jr, 2011). As before, the endogenous variable, an open house, is instrumented first. Then, the proportional hazard regression model with an instrumental variable (IV) design is estimated as follows (Tchetgen et al., 2015):

$$\begin{aligned} \ln\left(\frac{h(t|i,c)}{h_0(t|i,c)}\right) &= X_{ic}\Gamma + \mathbb{1}(\widehat{\text{open}})_{ic}\beta_1 + \text{unem}_c\beta_2 \\ \mathbb{1}(\text{open})_{ic} &= X_{ic}\Pi + \text{opencity}_c\lambda + \nu_{ic} \end{aligned}, \tag{3.6}$$

where $h(t|i,c)$ denotes the hazard function for the house i in the city c at time t. $h_0(t|i,c)$ denotes the baseline hazard function at time t where all regressors equal 0. The hazard ratio, $\frac{h(t|i,c)}{h_0(t|i,c)}$, can be regarded as the relative risk of the event, i.e. being sold, occurring at time t. All other variables are defined in the same way as before. The main Cox regression survival model is estimated using maximum likelihood estimation, and robust standard errors are clustered at the city level.

3.5.4 Extended Sales Price Model

This paper further comprehensively examines how two marketing strategies, list price signaling, and an open house, influence sale prices in an extended model. Since we cannot observe sales prices for houses that are still unsold at the end of the study period, there could exist sample selection bias in the outcome variable. It becomes an issue when housing and agent characteristics are quite distinct between houses sold and those that are still on the market. Therefore, if being sold is not selected randomly, the sample of houses sold is probably not representative of the entire housing market (Ford et al., 2005). Then, the presence of sample selection on whether a house is sold could bias the estimation of a sales price model.

To fix the potential bias, we adopt a Heckman two-step estimation procedure with a sample selection correction to estimate the extended sales price model. The design of the two-step Heckman procedure is presented as follows (Wooldridge, 2016). In the first stage, we estimate the probability that a house is sold from the Probit model using the data of all houses:

$$\Pr\left(\mathbb{1}(\text{sold}) = 1|\widetilde{Z}\right) = \Phi\left(\widetilde{Z}\widehat{\delta}\right), \tag{3.7}$$

where the vector, $\widetilde{Z}$, is the superset of all price determinants on the entire sample. Then, we use $\widehat{\delta}$ estimated in the above Probit model to compute the inverse Mill's ratios for each house i, $\lambda\left(\widetilde{Z}_i\widehat{\delta}\right)$, in the subsample of houses that are sold with $\lambda\left(\widetilde{Z}_i\widehat{\delta}\right) = \frac{\phi\left(\widetilde{Z}_i\widehat{\delta}\right)}{\Phi\left(\widetilde{Z}_i\widehat{\delta}\right)}$, where $\Phi(\cdot)$ and $\phi(\cdot)$ denote the probability density function (pdf) and cumulative distribution function (cdf) of the standard normal distribution.

In the second stage, after the inverse Mill's ratios are computed for each house, they are incorporated into the main sales price model below to get rid of potential sample selection bias in the sale. The sales price model aims to analyze all relevant price determinants, including Zestimate, asymmetric list price signaling effects, an open house, housing and agent characteristics, local unemployment rate, and short-run weather conditions during an open house, in addition to city and time fixed effects. Then, we fit the main extended sales price model where sales prices are regressed on all price determinants and estimated inverse Mill's ratios as follows:

$$\begin{aligned}\text{salep}_{icm} =& \mathbb{1}(\text{above})_i \times (\text{listp} - \text{Zestimate})_i \beta_1 + \mathbb{1}(\text{below})_i \times (\text{Zestimate} - \text{listp})_i \beta_2 \\ &+ \text{Zestimate}_i \beta_3 + \mathbb{1}(\text{open})_i \beta_4 + \mathbb{1}(\text{open})_i \times W_i \Pi + X_{ic}\Gamma + \text{unem}_{cm}\beta_5 \\ &+ \lambda\left(\widetilde{Z}_i\widehat{\delta}\right)\beta_6 + \delta_c + \lambda_m + \varepsilon_{icm}\end{aligned} \quad , \tag{3.8}$$

where the dependent variable, salep_{icm}, is the eventual transaction price for the house i in city c sold in month m. $\mathbb{1}(\text{open})_i$ is an indicator variable that equals 1 if an open house is offered in first 10 days on the market. The vector, W_i, denotes weather conditions during the open house, if

any, and its interaction with an open house further explores the influence of short-run weather conditions on sales prices. X_{ic} is the vector of housing and agent characteristics, as specified before. The inverse Mill's ratios, $\lambda\left(\widetilde{X}_i\widehat{\delta}\right)$, is added as the Heckman correction. δ_c and λ_m are city and month fixed effects, controlling for the unobserved factors that impact sales prices across places and months. ε_{icm} is the error term, clustered at both city and month level. Other variables are identified the same as the model (3.4) before.

3.6 Empirical Results

This section presents empirical results on how well Zestimates predict eventual sales prices and how two marketing strategies, i.e., list price signaling and an open house, influence sale prices and time on the market.

3.6.1 Accuracy of Zestimates and List Price Signaling

Before showing the estimation results, we implement some statistical tests to explore how Zestimates influence the list price set by a seller's agent. Table 3.2 presents the standard Pearson correlation coefficients and Spearman's rank correlation coefficients among three price variables. It is seen that, generally, Zestimates are accurate and informative due to their high correlation with list prices and sales prices, which suggests that they become a significant reference for an initial list price and eventual sales price. However, the Zestimate is slightly less accurate than an initial list in predicting an eventual transaction price. Specifically, compared to Zestimate, list prices are closer to sales prices, which implies that sellers hold some hidden information of a specific dwelling that is not available to Zillow and the public.

Table 3.3 presents the estimation results of the baseline models (3.2-3.4) regarding the accuracy of Zestimates and list price signaling effects. It can be seen that a house's Zestimate has a large influence on its sales price and picks up most information on sales prices since the

Table 3.2: Correlation coefficients among list prices, Zestimates, and sales prices

	Standard Pearson			Spearman's Rank		
	list price	Zestimate	sales price	list price	Zestimate	sales price
list price	1.0000			1.0000		
Zestimate	0.9859	1.0000		0.9823	1.0000	
	(0.0000)			(0.0000)		
sales price	0.9455	0.9125	1.0000	0.9431	0.9121	1.0000
	(0.0000)	(0.0000)		(0.0000)	(0.0000)	

Notes: The table presents the standard Pearson correlation coefficients and Spearman's rank correlation coefficients among the three prices. The significance levels are presented in parentheses. A correlation coefficient larger than 0.5 suggests a strong correlation.

slope coefficients are close to 1. In the baseline model 3.2, over 90% of the variations in sales prices can be explained by Zestimates. Based on the model 3.3, column (2) shows that about an extra 4% of variations in sales prices are explained by the signaling effects of list prices and the standard error of the regression substantially decreases, as opposed to the baseline model 3.2. For a one-dollar increase in the list price given the Zestimate, the resulting sales price is expected to increase by only \$0.16, significantly smaller than the previous estimates.[18] The finding provides the first empirical evidence suggesting that, due to the existence of Zestimate in the housing market, sellers' real estate agents have much less bargaining power than before.

The estimation results of the model 3.4 in column (2) reveal that setting a list price higher than its Zestimate helps boost the sales price, while a lower list price reduces the sales price, holding other variables equal. However, there exist asymmetric signaling effects of setting a list price, comparing the magnitudes of list price signaling between above and below its Zestimate, as shown in column (3). If the list price is set above Zestimate, a one-dollar increase in the list price can raise the sales price by \$0.15, while, if the list price is below Zestimate, a one-dollar decrease in the list price further lowers the sales price by \$0.18. It implies that list prices above the Zestimate are not informative of the house's quality as the list price set below the Zestimates.

[18] Asabere and Huffman (1993) in their early work estimated that, holding other factors equal, setting a list price 1% higher on average causes 0.89% premium in actual sales prices.

Table 3.3: Estimation results of accuracy of Zestimates and list price signaling

Dependent variable	Sale price				
Variables	(1)	(2)	(3)	(4)	(5)
Zestimate	0.9655***	0.9322***	0.9397***	0.9421***	0.9372***
	(0.0007)	(0.0005)	(0.0009)	(0.0004)	(0.0006)
list price - Zestimate		0.1639***			
		(0.0012)			
$\mathbb{1}(\text{above}) \times (\text{listp} - \text{Zestimate})$			0.1513***	0.1469***	0.1498***
			(0.0085)	(0.0076)	(0.0052)
$\mathbb{1}(\text{below}) \times (\text{Zestimate} - \text{listp})$			-0.1821***	-0.1987***	-0.1932***
			(0.0073)	(0.0067)	(0.0078)
$\mathbb{1}(\text{open})$				2,783.81***	2,821.98***
				(23.03)	(30.89)
realtor's rating					821.78***
					(21.32)
work experience					678.18***
					(18.07)
SE of the regression (in 10^5)	2.2191	1.6391	1.6123	1.6098	1.5902
# of observations	218,147	218,147	218,147	218,147	218,147
R^2	0.9019	0.9458	0.9478	0.9492	0.9515

Notes: This table presents the empirical results about how accurate Zestimates are in predicting the transaction prices, controlling for an open house and agent characteristics. The robust standard errors are presented in parentheses and clustered by city. *** p<0.01, ** p<0.05, * p<0.1

If the list price is set lower than the Zestimate, it basically shows to buyers that there might be some unfavorable attributes that are observed by Zillow. Another scenario is that the homeowner prefers a quick sale and thus sets a lower price to attract more potential bidders. In that sense, the buyers' agents would realize the larger bargaining power due to the desire for quick sales, which eventually lowers the transaction price. Column (4) presents the impact of an open house on the sales price, and, on average, a speedy open house can bring a price premium of nearly $2,800 on the eventual transaction price. Moreover, an agent with a higher rating and more work experience can help make a deal with a higher sales price.

3.6.2 Probability of Sale

To investigate what impacts the probability of sale, the two-stage Probit model 3.5 with an endogenous selection of an open house is estimated. Since the proportion of holding an open house in a city is taken as an instrumental variable for an open house, we first explore the geographical variations in selecting an open house by city. Figure 3.4 presents the histogram of percents of an open house across cities and its kernel density regression.[19] It shows that there exist large differences in the customs of holding open houses across cities in the United States. Some major coastal cities in California are very active in the residential housing market in terms of holding an open house, e.g., San Diego (34%) and West Hollywood (48%).

Table 3.4 presents the empirical results of the instrumental variable Probit (IV-Probit) analysis, where the dependent variable is whether the real property is sold over the study period. It is seen that, generally, the estimation results between the regular Probit and IV-Probit models are similar to each other, except for an open house and several exogenous covariates. It implies that the selection of an open house is endogenous and should be instrumented.[20] Empirical results

[19]The kernel density estimations are implemented with the Nadaraya-Watson algorithm using a Gaussian kernel and default optimal bandwidth (Chen et al., 2020).

[20]Since there is only one instrumental variable, the model is a just-identified and cannot be tested for the IV's exogeneity (Andrews, 2019).

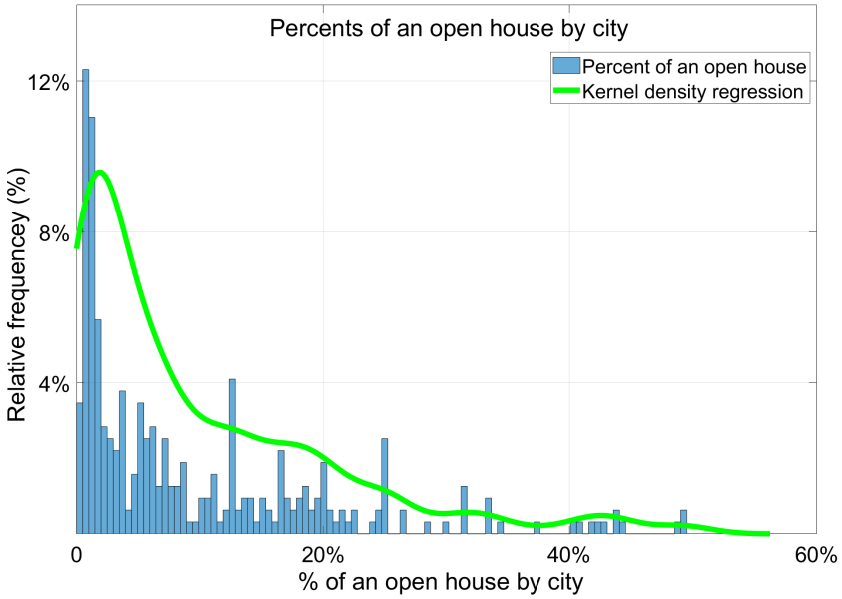


Figure 3.4: Geographical variations in percents of an open house across cities

of two relevant linear probability models are presented in Table 3.7 in Appendix.

The first-stage empirical results in Table 3.4 show that a speedy open house is more likely to be scheduled if the proportion of open houses in a local city is large and the list price of the house is set higher. Moreover, an open house occurs more often in a house with a bigger yard and more than one storey, while it is less likely to happen if there exists no cooling system and too many bedrooms. Apart from the housing attributes, a more experienced agent with a higher rating would more likely hold an open house.

In the second-stage estimation, we find that holding a speedy open house can increase the likelihood of selling a property, holding other factors equal. It is due to the fact that, during an open house, more potential buyers are expected to bid for the dwelling, which eventually expedites the selling process. Compared to its estimated value and the median sales price in the local city, a higher list price makes the property less likely to be sold. On top of that, there exists a nonlinear relationship between the price and the likelihood of being sold, where a higher list price disproportionately reduces the probability of a sale. In terms of housing attributes, a multi-storey house is less likely to be sold if it has more bedrooms than 90% of local houses and lacks a

Table 3.4: Probability of sale with a Probit model

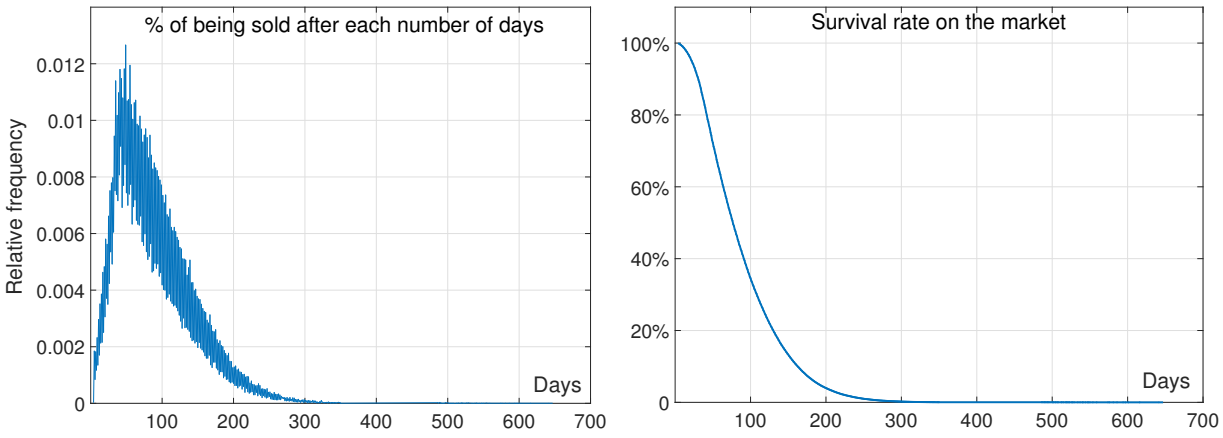
Model	Probit		IV-Probit	
	Coefficients	t-Stat.	Coefficients	t-Stat.
First-stage estimation				
Dependent variable: $\mathbb{1}(\text{open})$				
opencity			0.9882***	92.23
listoversalesp			0.0027	1.37
listoversalesp^2			-0.0011	-1.29
listoverZest			0.0064***	3.42
listoverZest^2			0.0003***	2.90
$\mathbb{1}(\text{manybed})$			-0.0062**	-2.28
cooling			0.0166**	1.97
yard size			0.0015***	10.15
multi-storey			0.0023**	2.11
realtor's rating			0.0212**	1.98
work experience			0.0121***	3.32
unemployment			-0.0008	-1.09
constant			0.0128***	4.59
Second-stage estimation				
Dependent variable: $\mathbb{1}(\text{sold})$				
$\widehat{\mathbb{1}(\text{open})}$	0.1027***	8.18	0.1438***	9.98
listoversalesp	-0.0232***	-5.38	-0.0185***	-3.39
listoversalesp^2	-0.0132***	-13.82	-0.0022***	-3.08
listoverZest	-0.0018***	-2.72	-0.0141***	-7.42
listoverZest^2	-0.0002***	-3.12	-0.0009***	-4.20
$\mathbb{1}(\text{manybed})$	-0.1172***	-23.98	-0.1115***	-18.09
cooling	0.0217***	3.31	0.0452***	15.98
yard size	0.0007**	2.03	0.0052***	9.15
multi-storey	-0.0078***	-3.09	-0.0092***	-4.31
realtor's rating	0.0127***	3.41	0.0211**	2.15
work experience	0.0021*	1.83	0.0110***	9.98
unemployment	-0.0157***	-8.92	-0.0128***	-11.91
constant	1.2157***	58.12	1.9828***	83.57
# of observations	220,218		220,218	
Pseudo R^2	0.0044		0.0048	

Notes: This table presents the empirical results of the Probit model and IV-Probit model 3.5. The robust standard errors are presented in the right columns of estimated coefficients and are clustered by city. *** p<0.01, ** p<0.05, * p<0.1

cooling system, holding other variables equal. For the agent characteristics, a house is more likely to be sold by an agent with a higher rating and more work experience. As an important location attribute, it is found that a higher unemployment rate in the local city significantly makes the residential housing market less active and decreases the chance of a successful deal.

3.6.3 Survival Analysis

Before discussing the estimation results of the survival model, we look at the selling time on the market and survival rate in Figure 3.5. Panel (a) presents the percents of houses being sold on each number of days after a dwelling is listed. Panel (b) shows the cumulative survival probability in the real estate market. It is be seen that most houses are sold nearly 40 days after being listed, and over 60% of the houses on the market are sold in the first 100 days.



(a) % of being sold across days (b) Cumulative survival probability

Figure 3.5: Time on the market and survival rate. This graph presents the percent of houses being sold after each number of days and cumulative survival probability on the market.

Following the time-on-market model (3.6), Table 3.5 presents the empirical results regarding how the selling time is influenced by a speedy open house in the first ten days, the listing price strategy, housing and agent characteristics, and the local unemployment rate. It is seen that holding a speedy open house in the first ten days, on average, increases the expected likelihood of sales by 17% relative to those without an open house. In terms of the list price strategy, empirical

Table 3.5: Estimation results of Cox regression model

Variable	Parameter	Hazard ratio	z-Stat.
$\mathbb{1}(\widehat{\text{open}})$	0.0682***	1.1701	5.07
listoversalesp	-0.0331***	0.9267	-4.79
listoversalesp^2	-0.0362***	0.9201	-3.21
listoverZest	-0.0408***	0.9102	-4.98
listoverZest^2	-0.0187***	0.9578	-3.01
$\mathbb{1}(\text{manybed})$	-0.0282***	0.9371	-4.15
cooling	0.0091***	1.0211	3.38
yard size	0.0133***	1.0312	4.57
multi-storey	-0.0232***	0.9479	-9.01
realtor's rating	0.0035***	1.0081	5.68
work experience	0.0014***	1.0032	5.19
unemployment	-0.0047***	0.9892	-12.81
$\mathbb{1}(\text{open})$ is instrumented in the first stage			
# of subjects	220,218		
# of failures	213,933		
Time at risk	19,499,583		
# of observations	220,218		
Pseudo R^2	0.0039		

Notes: This table presents the empirical results of the Cox regression model 3.6 with the instrumental variable design. The hazard ratios are the exponent of estimated coefficients. The robust standard errors are clustered by city and presented in the right columns of hazard ratios. *** p<0.01, ** p<0.05, * p<0.1

results suggest that houses whose list prices are set higher have a lower hazard of getting sold than others whose list prices are lower, relative to their local median sales prices and Zestimates. Having more bedrooms than 90% of local houses in a house is associated with a 7.3% decrease in the expected hazard of getting sold, while the equipment of a cooling system increases the expected hazard of sales by 2.9% compared with those without it. The expected hazard of one house to be sold is 5.2% lower if there exist multiple storeys, *ceteris paribus*. In addition, a house is more likely to be sold at any time if it has a bigger yard. As for agent characteristics, a more experienced seller's agent with a higher rating expedites the sales of residential real properties, *ceteris paribus*. At last, the probability of sales is lower in a local city with a higher unemployment rate, holding other factors equal.

3.6.4 Extended Sales Price Model

Following the Heckman two-step estimation procedure that corrects for sample selection bias, Table 3.6 presents the empirical results of the extended sales price model (3.8). It is seen that the coefficients on inverse Mill's ratios are statistically significant, suggesting the existence of sample selection bias and necessity for correcting the bias of sale in the sales price model.

Overall, the estimation results are robust to alternative model specifications, primarily due to the availability of Zestimates. Given the influence of Zestimates, housing attributes, location attributes, and time fixed effects have much less impact on the eventual transaction prices. Generally, the property values estimated by Zillow, Zestimates, have a statistically significant influence on sales prices at the 1% level. It implies that Zestimates give an important reference to an eventual transaction price. Both buyers and sellers of a real property take the estimates as an informative benchmark in the transactions. Moreover, most information has already been absorbed into the values estimated by Zillow using a complex hedonic pricing model that considers the time-varying market environment in neighboring areas.

Table 3.6 also suggests that the signaling strategy of a list price plays an important role

Table 3.6: Empirical results of the extended sales price model

Dependent variable	Sales price, salep_{icm}			
Variables	(1)	(2)	(3)	(4)
Zestimate	0.9245***	0.9202***	0.9271***	0.9311***
	(0.0010)	(0.0011)	(0.0009)	(0.0008)
$\mathbb{1}(\text{above}) \times (\text{listp} - \text{Zestimate})$	0.1421***	0.1413***	0.1432***	0.1312***
	(0.0071)	(0.0072)	(0.0072)	(0.0072)
$\mathbb{1}(\text{below}) \times (\text{Zestimate} - \text{listp})$	-0.1832***	-0.1798***	-0.1729***	-0.1817***
	(0.0088)	(0.0078)	(0.0083)	(0.0086)
$\mathbb{1}(\text{open})$	2,813.25***	2,922.18***	2,831.22***	2,839.89***
	(33.21)	(34.32)	(32.71)	(38.09)
$\mathbb{1}(\text{open}) \times$weather conditions				
temperature		0.0003	-0.0002	-0.0003
		(0.0042)	(0.0041)	(0.0052)
precipitation		-0.0017	-0.0018	-0.0011
		(0.0033)	(0.0039)	(0.0041)
snow		-0.0072***	-0.0087***	-0.0093***
		(0.0009)	(0.0010)	(0.0012)
wind		0.0101	-0.0083	0.0129
		(0.0243)	(0.0146)	(0.1235)
sun		0.0091**	0.0098**	0.0109***
		(0.0045)	(0.0042)	(0.0033)
realtor rating	672.31***	621.23***	574.13***	471.89***
	(25.12)	(29.24)	(32.21)	(26.09)
work experience	621.09***	631.98***	542.87***	592.62***
	(20.23)	(21.09)	(26.98)	(22.99)
unemployment	-0.0041**	-0.0052**	-0.0056**	-0.0048***
	(0.0003)	(0.0005)	(0.0004)	(0.0004)
inverse Mill's ratio	-0.0211**	-0.0312***	-0.0388***	-0.0392***
	(0.0042)	(0.0045)	(0.0046)	(0.0043)
Housing attributes	Y	Y	N	Y
City FE	N	N	Y	Y
Month FE	N	N	Y	Y
# of observations	218,147	218,147	218,147	218,147
SE of the regression (in 10^5)	1.4678	1.3921	1.3121	1.2723
R^2	0.9787	0.9798	0.9821	0.9868

Notes: This table presents the empirical results of the extended sales price model (3.8) with a Heckman correction. The robust standard errors are presented in parentheses and are clustered by city and month. *** p<0.01, ** p<0.05, * p<0.1.

in sales prices, and there exist asymmetric signaling price effects. A listing price higher than its estimated value helps boost up the eventual transaction price, while a list price lower than its Zestimate shows less confidence of a seller and lowers the sales price in the transaction, holding other attributes equal. Holding a speedy open house, on average, increases the transaction price of a single-family house by around $2,900, i.e., nearby 0.63% of the sales price, *ceteris paribus*. During the time period of an open house, more comfortable weather helps increase an eventual transaction price.[21] Specifically, heavy snow during the first open house lowers the price, while a sunny day has a positive impact on the sales price. The daily mean temperature, precipitation rate, and wind speed barely have an influence on the sales price. Generally, there are two potential channels through which short-run weather conditions influence the sales price. The first possibility is that the weather could impact the desirability of certain housing attribute.[22] The other is that the number of home buyers going to the open house is larger due to the good weather. According to the general auction theory, holding housing attributes equal, a house tends to be sold at a higher price having more bidders (Milgrom and Weber, 1982). However, due to data limitations, the estimates only represent the overall effects of short-run weather conditions.

On top of the list price signaling and open house, the main characteristics of a seller's agent and local unemployment rate are controlled for in the extended sales price model (3.8). It is seen in Table 3.6 that the estimates are in line with conventional wisdom. A house is more likely to be sold at a higher price by an agent with a higher rating and more work experience in a more active labor market.

[21] Roth Tran (2019) also presents similar evidence, showing that negative weather shocks reduce sales in the U.S retail market.

[22] In good weather, potential home buyers might be willing to pay a higher price premium on the backyard.

3.7 Conclusion

The considerable size of the U.S. residential real estate market motivates us to comprehensively explore the structure of how this market works at the ground level. Due to the availability of the unique dataset involving Zestimates and accurate timing information on open houses, we are able to unveil a rich picture from our analysis on how the housing market has been fundamentally changed by them. Using the data on over 50,000 transaction records from Zillow, this paper analyzes how single-family homes in the U.S. are transacted and provides the first empirical evidence on Zestimate and the marketing strategy of an open house reflected in sales performance of single-family houses.

Firstly, we examine whether the availability of Zestimate, the new publicly available price information, affects both listing and sales prices. Empirical results suggest that Zestimates play an important and complex role in driving the sales process. It is found that Zestimates deliver an important reference to an eventual transaction price and substantially reduce the information asymmetry in modern residential real estate markets. A lower listing price relative to its Zestimate expedites the sales of residential real properties and raises the likelihood of a real property being sold at an over-the-list price.

Apart from Zestimate, we then explore what influences a selling agent's decision to hold an open house and the role such an open house has on the sales price and sales timing. We find that an experienced seller's realtor is more likely to show the houses with preferred features during an open house. Holding an open house also substantially increases sales price, decreases time on the market, and brings a price premium. We also look at the exogenous weather realizations in the short periods when open houses are scheduled, and the findings show that more comfortable weather helps improve the sales performance of houses.

These findings have profound implications for homeowners and real estate agents in the modern housing market. For homeowners, they have easy access to a reference price, Zestimate,

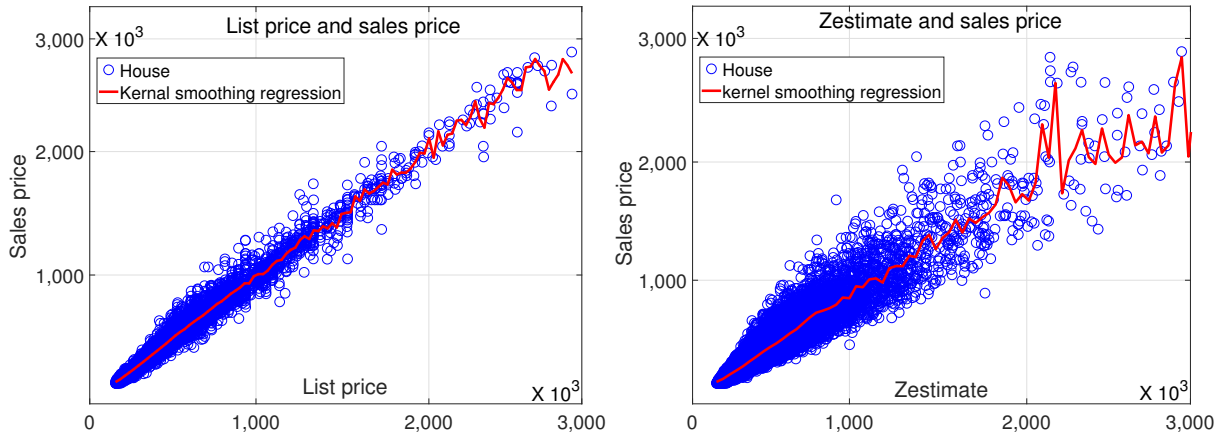
for the dwelling to be sold, which helps them hold a reasonable expectation on how high the sales price would be. Due to publicly available information, selling agents have less negotiation power and could hide less information on the dwelling characteristics, on the one hand. On the other hand, a buyer's agent attains an estimated price and more details on the home attributes, thus being able to take advantage of information completeness to find out a better deal. The analysis of an open house strategy is also of high interest to a selling agent when deciding on whether to hold an open house. Considering the relatively high cost of holding an open house to the seller's agent, the empirical results in this paper give much information on the potential benefits the marketing tool would bring, providing the economic rationale for holding an open house. Overall, the findings can be broadly adopted to make wiser decisions for participants in the residential real estate market.

3.8 Acknowledgement

Chapter 3, in full, is a coauthored work with Richard T. Carson. Richard T. Carson; Jiajun Lu. It is currently in preparation for publication and the dissertation author was the primary researcher of this chapter.

3.9 Appendix

3.9.1 Supplemental Table and Graph



(a) List price and sales price

(b) Zestimate and sales price

Figure 3.6: Kernel smoothing regression plot of list price, sales price, and Zestimate. The graphs illustrate the bivariate kernel regression plots of sales price on Zestimate and sales price on list price, respectively. The kernel density estimations are implemented with Nadaraya-Watson algorithm using a Gaussian kernel and default optimal bandwidth. Prices are measured in 2020 U.S. dollar.

Table 3.7: Probability of sale with a linear probability model

Model	LPM		IV-LPM	
	Coefficients	t-Stat.	Coefficients	t-Stat.
Second-stage estimation				
Dependent variable: $\mathbb{1}(\text{sold})$				
$\mathbb{1}(\widehat{\text{open}})$	0.1201***	7.98	0.1311***	8.87
listoversalesp	-0.0123***	-6.81	-0.0134***	-4.92
listoversalesp^2	-0.0008***	-9.13	-0.0011***	-8.12
listoverZest	-0.0012***	-4.23	-0.0014***	-3.41
listoverZest^2	-0.0003***	-7.22	-0.0004***	-6.21
$\mathbb{1}(\text{manybed})$	-0.1421***	-33.11	-0.1653***	-31.92
cooling	0.0311***	4.32	0.0322***	5.28
yard size	0.0003***	2.98	0.0005***	2.45
multi-storey	-0.0122***	-4.21	-0.0145***	-3.19
realtor's rating	0.0412***	4.11	0.0371***	3.15
work experience	0.0231**	2.01	0.0209***	2.38
unemployment	-0.0005*	-1.72	-0.0008**	-2.11
constant	0.0357***	2.21	0.0481***	2.71
# of observations	220,218		220,218	
R^2	0.0014		0.0015	

Notes: This table presents empirical results of the linear probability model (LPM) and IV-LPM with the same variables as in the model 3.5. The first-stage empirical results are omitted here since the equation with the instrumented variable is the same as in Table 3.4. The robust standard errors are presented in the right columns of estimated coefficients and are clustered by city. *** p<0.01, ** p<0.05, * p<0.1

Bibliography

[1] John Abraham and J Hunt. Specification and estimation of nested logit model of home, workplaces, and commuter mode choices by multiple-worker households. *Transportation Research Record: Journal of the Transportation Research Board*, 3(1606):17–24, 1997.

[2] Ather H Akbari and Yigit Aydede. Effects of immigration on house prices in Canada. *Applied Economics*, 44(13):1645–1658, 2012.

[3] David Albouy. Are big cities really bad places to live? Improving quality-of-life estimates across cities. *National Bureau of Economic Research*, 14472:41220–48109, 2008.

[4] David Albouy. What are cities worth? Land rents, local productivity, and the total value of amenities. *Review of Economics and Statistics*, 98(3):477–487, 2016.

[5] David Albouy, Walter Graf, Ryan Kellogg, and Hendrik Wolff. Climate amenities, climate change, and American quality of life. *Journal of the Association of Environmental and Resource Economists*, 3(1):205–246, 2016.

[6] Marcus T Allen, Anjelita Cadena, Jessica Rutherford, and Ronald C Rutherford. Effects of real estate brokers' marketing strategies: Public open houses, broker open houses, MLS virtual tours, and MLS photographs. *Journal of Real Estate Research*, 37(3):343–369, 2015.

[7] William Alonso. A theory of the urban land market. *Papers in Regional Science*, 6(1): 149–157, 1960.

[8] Alex Anas. *Residential Location Markets and Urban Transportation. Economic Theory, Econometrics and Policy Analysis With Discrete Choice Models*. New York Academic Press, 1982.

[9] Alex Anas. Discrete choice theory, information theory and the multinomial logit and gravity models. *Transportation Research Part B: Methodological*, 17(1):13–23, 1983.

[10] Isaiah Andrews. On the structure of IV estimands. *Journal of econometrics*, 211(1): 294–307, 2019.

[11] Luc Anselin. Lagrange multiplier test diagnostics for spatial dependence and spatial heterogeneity. *Geographical Analysis*, 20(1):1–17, 1988.

[12] Luc Anselin and Nancy Lozano-Gracia. Errors in variables and spatial effects in hedonic house price models of ambient air quality. *Empirical Economics*, 34:5–34, 2008.

[13] Paul K Asabere and Forrest E Huffman. Price concessions, time on the market, and the actual sale price of homes. *Journal of Real Estate Finance and Economics*, 6(2):167–174, 1993.

[14] Alan Barreca, Karen Clay, Olivier Deschênes, Michael Greenstone, and Joseph S Shapiro. Convergence in adaptation to climate change: Evidence from high temperatures and mortality, 1900-2004. *American Economic Review*, 105(5):247–251, 2015.

[15] Robert J Barro, Xavier Sala-i Martin, Olivier Jean Blanchard, and Robert E Hall. Convergence across states and regions. *Brookings Papers on Economic Activity*, pages 107–182, 1991.

[16] Patrick Bayer and Christopher Timmins. Estimating equilibrium models of sorting across locations. *The Economic Journal*, 117(518):353–374, 2007.

[17] Patrick Bayer, Nathaniel Keohane, and Christopher Timmins. Migration and hedonic valuation: The case of air quality. *Journal of Environmental Economics and Management*, 58(1):1–14, 2009.

[18] Patrick Bayer, Robert McMillan, Alvin Murphy, and Christopher Timmins. A dynamic model of demand for houses and neighborhoods. *Econometrica*, 84(3):893–942, 2016.

[19] Ali Behnood, Mahsa Modiri-Gharehveran, and Arash Moradkhani Roshandeh. The effects of drivers' behavior on driver-injury severities in Iran: An application of the mixed-logit model. *Scientia Iranica*, 23(6):2429–2440, 2016.

[20] Moshe E Ben-Akiva, Steven R Lerman, and Steven R Lerman. *Discrete choice analysis: theory and application to travel demand*, volume 9. MIT press, 1985.

[21] Roland Benabou. Workings of a city: location, education, and production. *The Quarterly Journal of Economics*, 108(3):619–652, 1993.

[22] Chandra R Bhat and Rachel Gossen. A mixed multinomial logit model analysis of weekend recreational episode type choice. *Transportation Research Part B: Methodological*, 38(9): 767–787, 2004.

[23] Michel Bierlaire. BIOGEME: a free package for the estimation of discrete choice models. In *Swiss Transport Research Conference*, pages 32–24, 2003.

[24] George J Borjas. Native internal migration and the labor market impact of immigration. *Journal of Human Resources*, 41(2):221–258, 2006.

[25] Leah Platt Boustan, Price V Fishback, and Shawn Kantor. The effect of internal migration on local labor markets: American cities during the Great Depression. *Journal of Labor Economics*, 28(4):719–746, 2010.

[26] Regina M Bures. Migration and the life course: is there a retirement transition? *International Journal of Population Geography*, 3(2):109–119, 1997.

[27] Meghan R Busse, Devin G Pope, Jaren C Pope, and Jorge Silva-Risso. The psychological effect of weather on car purchases. *Quarterly Journal of Economics*, 130(1):371–414, 2015.

[28] Van Butsic, Ellen Hanak, and Robert G Valletta. Climate change and housing prices: Hedonic estimates for ski resorts in western North America. *Land Economics*, 87(1): 75–91, 2011.

[29] Gavin Cameron, John Muellbauer, and Anthony Murphy. Housing market dynamics and regional migration in Britain. In *CEPR Discussion Paper No. 5832*, pages 1–57. CEPR discussion paper, 2006.

[30] Simone Caschili and Andrea De Montis. Accessibility and complex network analysis of the US commuting system. *Cities*, 30:4–17, 2013.

[31] Robert Cervero. Public transport and sustainable urbanism: global lessons. In *Transit Oriented Development*, pages 43–56. Routledge, 2016.

[32] Emmanouil Chaniotakis and Adam J Pel. Drivers' parking location choice under uncertain parking availability and search times: A stated preference experiment. *Transportation Research Part A: Policy and Practice*, 82:228–239, 2015.

[33] Junhua Chen, Fei Guo, and Ying Wu. One decade of urban housing reform in China: Urban housing price dynamics and the role of migration and urbanization, 1995–2005. *Habitat International*, 35(1):1–8, 2011.

[34] Xin Chen, Xuejun Ma, and Wang Zhou. Kernel density regression. *Journal of Statistical Planning and Inference*, 205:318–329, 2020.

[35] Yong Chen and Stuart S Rosenthal. Local amenities and life-cycle migration: Do people move for jobs or fun? *Journal of Urban Economics*, 64(3):519–537, 2008.

[36] Zhenhua Chen and Kingsley E Haynes. Impact of high speed rail on housing values: An observation from the Beijing–Shanghai line. *Journal of Transport Geography*, 43:91–100, 2015.

[37] David E Clark and William J Hunter. The impact of economic opportunity, amenities and fiscal factors on age-specific migration rates. *Journal of Regional Science*, 32(3):349–365, 1992.

[38] Karen Schmith Conway and Andrew J Houtenville. Do the elderly "vote with their feet?". *Public Choice*, 97(4):663–685, 1998.

[39] Charles Corcoran and Fei Liu. Accuracy of Zillow's Home Value Estimates. *Real Estate Issues*, 39(1):45–49, 2014.

[40] Andrew Daly and Stanley Zachary. Improved multiple choice models. *Determinants of Travel Choice*, 335:357, 1978.

[41] Paul S Davies, Michael J Greenwood, and Haizheng Li. A conditional logit approach to US state-to-state migration. *Journal of Regional Science*, 41(2):337–360, 2001.

[42] Jakob De Haan. Political institutions and economic growth reconsidered. *Public Choice*, 131(3-4):281–292, 2007.

[43] Olivier Deschênes and Michael Greenstone. Climate change, mortality, and adaptation: Evidence from annual fluctuations in weather in the US. *American Economic Journal: Applied Economics*, 3(4):152–185, 2011.

[44] Sanjaya DeSilva, Anh Pham, and Michael Smith. Racial and ethnic price differentials in a small urban housing market. *Housing Policy Debate*, 22(2):241–269, 2012.

[45] Rebecca Diamond. The determinants and welfare implications of us workers' diverging location choices by skill: 1980-2000. *American Economic Review*, 106(3):479–524, 2016.

[46] Rebecca Diamond, Tim McQuade, and Franklin Qian. The effects of rent control expansion on tenants, landlords, and inequality: Evidence from San Francisco. *American Economic Review*, 109(9):3365–3394, 2019.

[47] Yi Dou, Xiao Luo, Liang Dong, Chuntao Wu, Hanwei Liang, and Jingzheng Ren. An empirical study on transit-oriented low-carbon urban land use planning: Exploratory Spatial Data Analysis (ESDA) on Shanghai, China. *Habitat International*, 53:379–389, 2016.

[48] William Duncombe, Mark Robbins, and Douglas Wolf. *Chasing the Elderly: Can State and Local Governments Attract Recent Retirees?* Center for Policy Research, 2000.

[49] William Duncombe, Mark Robbins, and Douglas A Wolf. Retire to where? A discrete choice model of residential location. *International Journal of Population Geography*, 7(4): 281–293, 2001.

[50] Gilles Duranton and Matthew A Turner. Urban form and driving: Evidence from US cities. *Journal of Urban Economics*, 108:170–191, 2018.

[51] Imke Durre, Matthew J Menne, Byron E Gleason, Tamara G Houston, and Russell S Vose. Comprehensive automated quality assurance of daily surface observations. *Journal of Applied Meteorology and Climatology*, 49(8):1615–1633, 2010.

[52] J Paul Elhorst. Applied spatial econometrics: raising the bar. *Spatial Economic Analysis*, 5(1):9–28, 2010.

[53] Glenn Ellison, Edward L Glaeser, and William R Kerr. What causes industry agglomeration? Evidence from coagglomeration patterns. *American Economic Review*, 100(3): 1195–1213, 2010.

[54] Qin Fan, H Allen Klaiber, and Karen Fisher-Vanden. Does extreme weather drive interregional brain drain in the US? Evidence from a sorting model. *Land Economics*, 92(2): 363–388, 2016.

[55] Qin Fan, Karen Fisher-Vanden, and H Allen Klaiber. Climate Change, Migration, and Regional Economic Impacts in the United States. *Journal of the Association of Environmental and Resource Economists*, 5(3):643–671, 2018.

[56] Gema Fernández-Avilés, Roman Minguez, and José-Mar\'\ia Montero. Geostatistical air pollution indexes in spatial hedonic models: the case of Madrid, Spain. *Journal of Real Estate Research*, 34(2):243–274, 2012.

[57] James Scott Ford, Ronald C Rutherford, and Abdullah Yavas. The effects of the internet on marketing residential real estate. *Journal of Housing Economics*, 14(2):92–108, 2005.

[58] Amnon Frenkel and Sigal Kaplan. The joint choice of tenure, dwelling type, size and location: the effect of home-oriented versus culture-oriented lifestyle. *Letters in Spatial and Resource Sciences*, 8(3):233–251, 2015.

[59] Elaine F Frey, Marissa B Palin, Patrick J Walsh, and Christine R Whitcraft. Spatial hedonic valuation of a multi-use urban wetland in southern California. *Agricultural and Resource Economics Review*, 42(2):387–402, 2013.

[60] Milton Friedman. The permanent income hypothesis. In *A Theory of the Consumption Function*, pages 20–37. Princeton University Press, 1957.

[61] Salim Furth and Olivia Gonzalez. California Zoning: Housing Construction and a New Ranking of Local Land Use Regulation. *Mercatus Research Paper Forthcoming*, 2019.

[62] Victor Gan, Vaishali Agarwal, and Ben Kim. Data Mining Analysis and Predictions of Real Estate Prices. *Issues in Information System*, 16:30–36, 2015.

[63] Peter Ganong and Daniel Shoag. Why has regional income convergence in the US declined? *Journal of Urban Economics*, 102:76–90, 2017.

[64] Edward L Glaeser and Bryce A Ward. The causes and consequences of land use regulation: Evidence from Greater Boston. *Journal of Urban Economics*, 65(3):265–278, 2009.

[65] Edward L Glaeser, Joseph Gyourko, and Raven E Saks. Why have housing prices gone up? *American Economic Review*, 95(2):329–333, 2005.

[66] Allen C Goodman and Thomas G Thibodeau. Housing market segmentation. *Journal of Housing Economics*, 7(2):121–143, 1998.

[67] Richard K Green and Hyojung Lee. Age, demographics, and the demand for housing, revisited. *Regional Science and Urban Economics*, 61:86–98, 2016.

[68] Michael J Greenwood, Gary L Hunt, Dan S Rickman, and George I Treyz. Migration, regional equilibrium, and the estimation of compensating differentials. *The American Economic Review*, 81(5):1382–1390, 1991.

[69] Jeffrey Grogger and Gordon H Hanson. Income maximization and the selection and sorting of international migrants. *Journal of Development Economics*, 95(1):42–57, 2011.

[70] C Angelo Guevara and Moshe E Ben-Akiva. Sampling of alternatives in multivariate extreme value (MEV) models. *Transportation Research Part B: Methodological*, 48:31–52, 2013.

[71] Joseph Gyourko and Joseph Tracy. The structure of local public finance and the quality of life. *Journal of Political Economy*, 99(4):774–806, 1991.

[72] Joseph Gyourko, Christopher Mayer, and Todd Sinai. Superstar cities. *American Economic Journal: Economic Policy*, 5(4):167–199, 2013.

[73] James R Hagerty. How Good Are Zillow's Estimates? *Wall Street Journal*, 2(14):476–478, feb 2007.

[74] Lu Han and William C Strange. What is the role of the asking price for a house? *Journal of Urban Economics*, 93:115–130, 2016.

[75] Keith Head, John Ries, and Deborah Swenson. Agglomeration benefits and location choice: Evidence from Japanese manufacturing investments in the United States. *Journal of International Economics*, 38(3-4):223–247, 1995.

[76] David A Hensher and Lester W Johnson. *Applied discrete-choice modelling*. Routledge, 2018.

[77] David A Hensher, John M Rose, and William H Greene. *Applied choice analysis: a primer*. Cambridge University Press, 2005.

[78] Robert J Hill and Iqbal A Syed. Hedonic price–rent ratios, user cost, and departures from equilibrium in the housing market. *Regional Science and Urban Economics*, 56:60–72, 2016.

[79] Chinh Ho and David Hensher. Housing prices and price endogeneity in tenure and dwelling type choice models. *Case Studies on Transport Policy*, 2(3):107–115, 2014.

[80] Arne Risa Hole. Fitting mixed logit models by using maximum simulated likelihood. *The Stata Journal*, 7(3):388–401, 2007.

[81] Joel L Horowitz. The role of the list price in housing markets: theory and an econometric model. *Journal of Applied Econometrics*, 7(2):115–129, 1992.

[82] Biqing Huang and Ronald Rutherford. Who you going to call? Performance of realtors and non-realtors in a MLS setting. *Journal of Real Estate Finance and Economics*, 35(1): 77–93, 2007.

[83] P Wilner Jeanty, Mark Partridge, and Elena Irwin. Estimation of a spatial simultaneous equation model of population migration and housing price dynamics. *Regional Science and Urban Economics*, 40(5):343–352, 2010.

[84] Jangik Jin and Hee-Yeon Lee. Understanding residential location choices: an application of the UrbanSim residential location model on Suwon, Korea. *International Journal of Urban Sciences*, 22(2):216–235, 2018.

[85] William R Johnson. House prices and female labor force participation. *Journal of Urban Economics*, 82:1–11, 2014.

[86] Thomas R Karl, Jerry M Melillo, Thomas C Peterson, and Susan J Hassol. *Global climate change impacts in the United States*. New York: Cambridge University Press, 2009.

[87] Harry H Kelejian and Ingmar R Prucha. HAC estimation in a spatial framework. *Journal of Econometrics*, 140(1):131–154, 2007.

[88] Thomas A Knapp, Nancy E White, and David E Clark. A nested logit approach to household mobility. *Journal of Regional Science*, 41(1):1–22, 2001.

[89] Sriram Krishnamurthy and Kara Maria Kockelman. Propagation of uncertainty in transportation land use models: Investigation of dram-empal and utpp predictions in austin, texas. *Transportation Research Record*, 1831(1):219–229, 2003.

[90] Douglass B Lee. Retrospective on large-scale urban models. *Journal of the American*

Planning Association, 60(1):35–40, 1994.

[91] Hyojung Lee. Are Millennials Coming to Town? Residential Location Choice of Young Adults. *Urban Affairs Review*, pages 1–40, 2018.

[92] Kwan Ok Lee and Gary Painter. Housing tenure transitions of older households: What is the role of child proximity? *Real Estate Economics*, 42(1):109–152, 2014.

[93] James LeSage and Robert Kelley Pace. *Introduction to spatial econometrics*. Chapman and Hall/CRC, 2009.

[94] Kao-Lee Liaw and Jacques Ledent. Nested logit model and maximum quasi-likelihood method: A flexible methodology for analyzing interregional migration patterns. *Regional Science and Urban Economics*, 17(1):67–88, 1987.

[95] Charles F Longino. *Retirement migration in America: An analysis of the size, trends and economic impact of the country's newest growth industry*. Vacation Place, 1995.

[96] Ira S Lowry. A Model of Metropolis. Technical report, RAND CORP SANTA MONICA CALIF, 1964.

[97] Stephen Malpezzi. Hedonic pricing models: a selective and applied review. *Housing Economics and Public Policy*, 2(5):67–89, 2002.

[98] Daniel McFadden. Conditional logit analysis of qualitative choice behavior. *Frontier in Econometrics*, 1973.

[99] Daniel McFadden. Modeling the choice of residential location. *Transportation Research Record*, 1(673):72–77, 1978.

[100] Daniel McFadden and Kenneth Train. Mixed MNL models for discrete response. *Journal of Applied Econometrics*, 15(5):447–470, 2000.

[101] Stanley McGreal, Alastair Adair, Louise Brown, and James Webb. Pricing and time on the market for residential properties in a major UK city. *Journal of Real Estate Research*, 31 (2):209–233, 2009.

[102] Paul R Milgrom and Robert J Weber. A theory of auctions and competitive bidding. *Econometrica*, 50(5):1089–1122, 1982.

[103] Rupert G Miller Jr. *Survival analysis*, volume 66. John Wiley & Sons, 2011.

[104] Johan A Mistiaen and Ivar E Strand. Location choice of commercial fishermen with heterogeneous risk preferences. *American Journal of Agricultural Economics*, 82(5): 1184–1190, 2000.

[105] Raven Molloy, Christopher L Smith, and Abigail Wozniak. Internal migration in the United States. *Journal of Economic Perspectives*, 25(3):173–196, 2011.

[106] Enrico Moretti. *The new geography of jobs*. Longman London, 2012.

[107] Jennifer Murdock. Handling unobserved site characteristics in random utility models of recreation demand. *Journal of Environmental Economics and Management*, 51(1):1–25, 2006.

[108] Abeba Mussa, Uwaoma G Nwaogu, and Susan Pozo. Immigration and housing: A spatial econometric analysis. *Journal of Housing Economics*, 35:13–25, 2017.

[109] L Rachel Ngai and Silvana Tenreyro. Hot and cold seasons in the housing market. *American Economic Review*, 104(12):3991–4026, 2014.

[110] Guy H Orcutt. A new type of socio-economic system. *The Review of Economics and Statistics*, pages 116–123, 1957.

[111] Liv Osland and Inge Thorsen. Spatial impacts, local labour market characteristics and housing prices. *Urban Studies*, 50(10):2063–2083, 2013.

[112] Gianmarco I P Ottaviano and Giovanni Peri. Rethinking the effect of immigration on wages. *Journal of the European Economic Association*, 10(1):152–197, 2012.

[113] Gianmarco I P Ottaviano, Giovanni Peri, and Others. The effects of immigration on US wages and rents: A general equilibrium approach. *Migration Impact Assessment: New Horizons*, pages 107–146, 2012.

[114] Francesca Pagliara, John Preston, and David Simmonds. *Residential location choice: Models and applications*. Springer Science & Business Media, 2010.

[115] Monika Piazzesi, Martin Schneider, and Johannes Stroebel. Segmented housing search. *American Economic Review*, 110(3):720–759, 2020.

[116] Andrew J Plantinga, Cécile Détang-Dessendre, Gary L Hunt, and Virginie Piguet. Housing prices and inter-urban migration. *Regional Science and Urban Economics*, 43(2):296–306, 2013.

[117] Dudley L Poston, Li Zhang, David J Gotcher, and Yuan Gu. The effect of climate on migration: United States, 1995–2000. *Social Science Research*, 38(3):743–753, 2009.

[118] Diego Puga. The magnitude and causes of agglomeration economies. *Journal of Regional Science*, 50(1):203–219, 2010.

[119] Stephen H Putman. DRAM/EMPAL ITLUP. *Integrated Transportation Land-Use Activity*

Allocation Models: General Description. SH Putman Associates, 1991.

[120] Stephen H Putman and Shih-Liang Chan. Planning Support System: Urban Models and GIS. *Planning support systems: integrating geographic information systems, models, and visualization tools*, page 99, 2001.

[121] Stephen H Putman and Frederick W Ducca. Calibrating urban residential models 2: empirical results. *Environment and Planning A*, 10(9):1001–1014, 1978.

[122] Birgitta Rabe and Mark Taylor. Residential mobility, quality of neighbourhood and life course events. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, 173 (3):531–555, 2010.

[123] Carol Rapaport. Housing demand and community choice: an empirical analysis. *Journal of Urban Economics*, 42(2):243–260, 1997.

[124] Spencer Rascoff and Stan Humphries. *Zillow Talk: Rewriting the Rules of Real Estate*. Grand Central Publishing, 2015.

[125] Katrin Rehdanz. Hedonic pricing of climate change impacts to households in Great Britain. *Climatic Change*, 74(4):413–434, 2006.

[126] David Revelt and Kenneth Train. Customer-specific taste parameters and mixed logit: Households' choice of electricity supplier. *University of California, Berkeley*, 2000.

[127] John Ries and Tsur Somerville. School quality and residential property values: evidence from Vancouver rezoning. *The Review of Economics and Statistics*, 92(4):928–944, 2010.

[128] Jennifer Roback. Wages, rents, and the quality of life. *Journal of Political Economy*, 90 (6):1257–1278, 1982.

[129] Sherwin Rosen. Hedonic prices and implicit markets: product differentiation in pure competition. *Journal of Political Economy*, 82(1):34–55, 1974.

[130] Sherwin Rosen. Wage-based indexes of urban quality of life. *Current Issues in Urban Economics*, pages 74–104, 1979.

[131] Brigitte Roth Tran. Long-and Short-Run Adaptation of Consumption to Weather and Climate Change. *Available at SSRN 3337110*, 2019.

[132] Ronald C Rutherford, Tom M Springer, and Abdullah Yavas. Conflicts between principals and agents: evidence from residential brokerage. *Journal of Financial Economics*, 76(3): 627–665, 2005.

[133] Patrick M Schirmer, Christof Zöllig, Kirill Müller, and Kay W Axhausen. Land use and

transport microsimulation on the canton of Zürich using UrbanSim. In *Integrated Transport and Land Use Modeling for Sustainable Cities*, pages 461–509. EPFL Press: distributed by Routledge, 2015.

[134] Wolfram Schlenker and Michael J Roberts. Nonlinear temperature effects indicate severe damages to US crop yields under climate change. *Proceedings of the National Academy of sciences*, 106(37):15594–15598, 2009.

[135] Mary Jo Schneider and Bernal L Green. A demographic and economic comparison of nonmetropolitan retirement and nonretirement counties in the US. *Journal of Applied Sociology*, 9:63–84, 1992.

[136] Uri Simonsohn. Weather to go to college. *Economic Journal*, 120(543):270–280, 2010.

[137] Paramita Sinha and Maureen L Cropper. The value of climate amenities: Evidence from US migration decisions. Technical report, National Bureau of Economic Research, 2013.

[138] Paramita Sinha, Martha L Caulkins, and Maureen L Cropper. Household location decisions and the value of climate amenities. *Journal of Environmental Economics and Management*, 2017.

[139] Paramita Sinha, Martha L Caulkins, and Maureen L Cropper. Household location decisions and the value of climate amenities. *Journal of Environmental Economics and Management*, 92:608–637, 2018.

[140] Kenneth A Small. A discrete choice model for ordered alternatives. *Econometrica*, 55(2): 409–424, 1987.

[141] Kim S So, Peter F Orazem, and Daniel M Otto. The effects of housing prices, wages, and commuting time on joint residential and job location choices. *American Journal of Agricultural Economics*, 83(4):1036–1048, 2001.

[142] Dan L. Swango. Resources for Real Estate Analysts and Valuers: Searchable Database Websites. *The Appraisal Journal*, 83(3):237, 2015.

[143] Karl E Taylor, Ronald J Stouffer, and Gerald A Meehl. An overview of CMIP5 and the experiment design. *Bulletin of the American Meteorological Society*, 93(4):485–498, 2012.

[144] Eric J Tchetgen Tchetgen, Stefan Walter, Stijn Vansteelandt, Torben Martinussen, and Maria Glymour. Instrumental variable estimation in a survival context. *Epidemiology (Cambridge, Mass.)*, 26(3):402, 2015.

[145] Jonathan Temple. Growth regressions and what the textbooks don't tell you. *Bulletin of Economic research*, 52(3):181–205, 2000.

[146] Peter Vovsha. *The cross-nested logit model: application to mode choice in the Tel-Aviv metropolitan area*. Transportation Research Board, 1997.

[147] Paul Waddell. A behavioral simulation model for metropolitan policy analysis and planning: residential location and housing market components of UrbanSim. *Environment and planning B: planning and design*, 27(2):247–263, 2000.

[148] Paul Waddell, Chandra Bhat, Naveen Eluru, Liming Wang, and Ram M Pendyala. Modeling interdependence in household residence and workplace choices. *Transportation Research Record*, 2003(1):84–92, 2007.

[149] Dantong Wang and Chun Yuan. Modeling and forecasting household energy consumption and related CO2 emissions integrating UrbanSim and transportation models: an Atlanta BeltLine case study. *Transportation Planning and Technology*, 41(4):448–462, 2018.

[150] Chieh-Hua Wen and Frank S Koppelman. The generalized nested logit model. *Transportation Research Part B: Methodological*, 35(7):627–641, 2001.

[151] William C Wheaton and Mark J Lewis. Urban wages and labor market agglomeration. *Journal of Urban Economics*, 51(3):542–562, 2002.

[152] Jeffrey M Wooldridge. *Introductory econometrics: A modern approach*. Nelson Education, 2016.

[153] Judith Yates and Daniel F Mackay. Discrete choice modelling of urban housing markets: A critical review and an application. *Urban Studies*, 43(3):559–581, 2006.

[154] Bin Zhou and Kara M Kockelman. Microsimulation of residential land development and household location choices: bidding for land in Austin, Texas. *Transportation Research Record*, 2077(1):106–112, 2008.