PREDICTING U.S. MIGRATION FLOWS WITH ONLINE SEARCH DATA

by

JAEYONG YOO

(Under the Direction of Patryk Babiarz)

ABSTRACT

High residential mobility is a distinguishing characteristic of American society. The majority of people move from one location to another at least once in their lives, and individual movement affects all geographic areas, urban and rural. Policymakers use government migration data to predict migration trends to implement or improve public policies. Because official migration data are released with a significant lag, it is difficult for policymakers to adjust housing and labor market policies for current and future migration trends. The Internet has become a default channel for information search about relocation and the beacon for estimating migration intentions. Because search engines provide real-time information, these online datasets can be exploited to predict current and future migration trends.

In this dissertation, I review internal migration in the United States over the past two decades and the determinants of the recent decline in migration rates. I then demonstrate how online search data can be used to measure migration intentions in origin states and help predict current and future interstate migration flows. I employ a combined panel data set of Internal Revenue Service (IRS) migration data and Google Trends data covering all 50 states and the District of Columbia. The data set contains a large set of macroeconomic indicators for each origin and destination state and migration flows between pairs of states. Using a gravity

approach with the balanced bilateral migration data, I show strong additional predictive power for interstate migration flows with Google Trends as compared to the baseline models that do not include online search data as covariates. My results imply that real-time online search data is effective in prognosticating actual moves and can be essential to managing and designing contingent migration policies.

INDEX WORDS: Migration, Residential Mobility, Prediction, Online Search, Big Data, Search Engine, Google Trends

PREDITING U.S. MIGRATION FLOWS WITH ONLINE SEARCH DATA

by

JAEYONG YOO

BA, Konkuk University, Republic of Korea, 2008

MS, Georgia State University, 2012

MS, Georgia State University, 2017

MA, Georgia State University, 2017

A Dissertation Submitted to the Graduate Faculty of The University of Georgia in Partial Fulfillment of the Requirements for the Degree

DOCTOR OF PHILOSOPHY

ATHENS, GEORGIA

2022

PREDICTING U.S. MIGRATION FLOWS WITH ONLINE SEARCH DATA

by

JAEYONG YOO

Major Professor:	Patryk Babiarz
Committee:	Velma Zahirovic-Herbert
	Swarn Chatterjee
	Sheri Worthy

Electronic Version Approved:

Ron Walcott
Vice Provost for Graduate Education and Dean of the Graduate School
The University of Georgia
May 2022

DEDICATION

To anyone who never gives up on their dream.

ACKNOWLEDGEMENTS

This dissertation would not have been accomplished without the help, support, and guidance of many people. First, I would like to thank my committee members, Drs. Patryk Babiarz, Velma Zahirovic-Herbert, Swarn Chatterjee, and Sheri Worthy. Especially, I am indebted to Dr. Patryk Babiarz, my major professor, who has looked after me in everything from my dissertation to the job, and Dr. Velma Herbert, my former co-chair and external member, who recruited me to the Ph.D. program and has continued to train me to become a researcher. I also appreciate the extraordinary support of Drs. Sophia Anong, Andrew Carswell, and Kimberly Skobba, Prof. Sherle Brown, and Melissa McBride in the Department of Financial Planning, Housing and Consumer Economics.

I would like to thank my family, Danah Jeong and Inna Yoo, for their incredible support, encouragement, and understanding. Their presence is the biggest motivation in my life. I also want to thank our parents and family in Korea and Australia. Especially, my parents, Byoungoh Yoo and Seoyeong Kang, have been the pillars of hope and support for me during my entire life. All credit goes to them.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

1.1 Background of Migration

High residential mobility is a distinguishing characteristic of American society. The majority of people move from one location to another at least once in their lives. There are two reasons why the study of residential mobility or migration is important.

First, migration is one of the most important factors in changes to urban and rural areas. Individuals and families move to fulfill their goals and desires, but their moves also contribute to economic, social, and demographic changes in a given area. Therefore, migration trends are of particular interest to policymakers. The study of migration can provide useful information about housing and labor market policies by adjusting those policies for current migration trends in a timely fashion. In some states, interstate migration is an important determinant of labor and housing market trends. For example, Frost (2020) documented that 71 percent of the growth in the number of Arizona's households in 2018 was due to domestic migration. High migration rates add flexibility to labor markets and alleviate income inequalities between areas. Moreover, migration correlates with demand for housing, which in turn affects housing and rental prices (Lin et al., 2018; Stawarz et al., 2021).

Second, understanding migration is important because migration involves many factors, such as economic, social, political, and environmental factors, and these determinants can be important sources for the evaluation of other decisions involving the public and private sectors (e.g., housing industry). To understand migration decisions, many scholars have been studying

the determinants of migration, documenting that the key determinants of migration include housing, job, and demographic factors, but many other factors also affect the migration decision.

The main reasons to move include housing (e.g., upgrading to better quality or a bigger house or new neighborhood), employment (e.g., new/lost job, or a shorter work commute), family (e.g., change in marital status) or starting a new chapter in life (e.g., relocating due to retirement). According to the Current Population Survey (CPS) data collected jointly by the U.S. Census Bureau and the U.S. Bureau of Labor Statistics (BLS), the within-county movers accounted for 59 percent of total movers between 2019 and 2020, whereas between-state movers accounted for only 14 percent during the same period. Most moves in the United States are short-distance (e.g., within-county moves), and short-distance movers tend to move for housing. Conversely, employment-related moves are more frequent among long-distance movers. The prevalence of long-distance moves (i.e., between-state movers) is much less frequent than short-distance moves; however, between-state moves (or interstate migration flows) are important because they might be a significant driver of economic growth.

Reliable projections of future migration rates provide local governments and policymakers with information necessary to plan fiscal policies and effective provision of government services. The growing interest in research on the determinants of interstate migration from scholars, policymakers, and practitioners, has led to a variety of attempts to assess and compare states' migration policies and governance in order to design policies that better attract migrants and/or retain residents (Frey, 2009).

The COVID-19 pandemic has impacted domestic migration patterns in the United States at the local and national levels (Frey, 2020). He pointed out that many people left large cities for suburbs or rural areas and many young adults returned to live with their other family members.

He also mentioned that the persistence of these pandemic-related migration patterns is unknown, which implies that these pandemic-related migration patterns can change at any time, depending on when the pandemic ends. Given this uncertainty, appropriate and accurate prediction of future migration flows is essential for policymakers.

Knowledge of migration data sources is important to policymakers and scholars for reliable projections of future migration rates. Three major sources of information on U.S. migration flows are the Current Population Survey (CPS), the American Community Survey (ACS), and the Internal Revenue Service (IRS). The CPS and ACS are the annual migration datasets released by the Census Bureau that provide broad migration statistics, while the IRS data provides a more detailed picture of annual migration flows by identifying both origins and destinations of moves. The data obtained from each of these sources differ in terms of periods for which the information is collected, demographic details, geographic granularity, and variables used to measure migration flows (Frey, 2009; Molloy et al., 2011).

The Current Population Survey is a national survey designed primarily to collect monthly information on labor force characteristics. The Current Population Survey sample size is about 60,000 households. The Annual Social and Economic Supplement to the Current Population Survey is conducted every year, usually in March. The Annual Social and Economic Supplement collects information on migration, as well as demographic and socioeconomic characteristics that are more detailed relative to the regular monthly Current Population Survey. The Annual Social and Economic Supplement sample size is about 99,000 households. The Current Population Survey has calculated migration rates since 1948, making it the longest publicly accessible secondary data on migration in the United States. The Current Population Survey provides

migration information at both regional and national levels, along with some measures of the reasons for moving and the distance of moves.

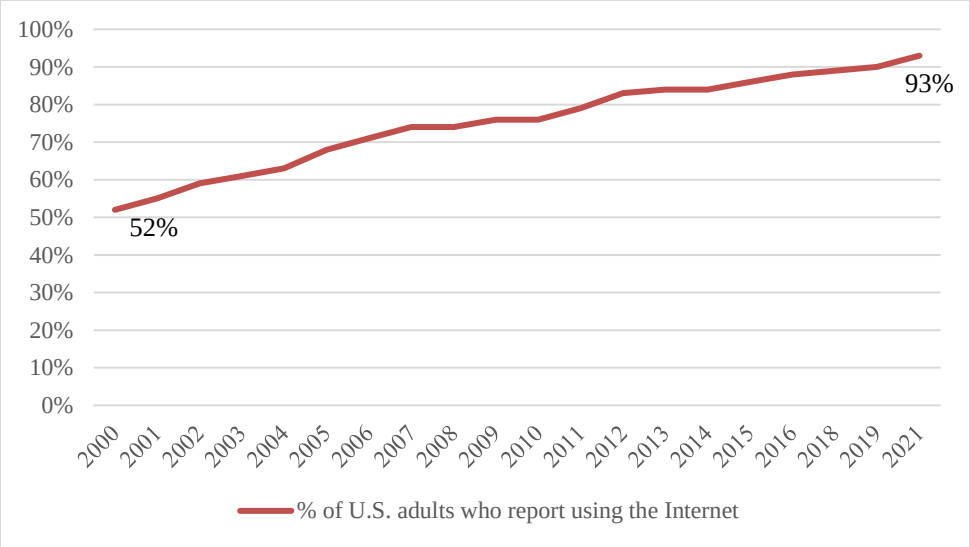
The American Community Survey is a primary source of detailed population and housing information. It provides information on annual county-to-county and state-to-state migration flows. The American Community Survey sample size is about 3.5 million housing units, which is considerably larger than that of the Current Population Survey. Since the American Community Survey also identifies geographic areas with much finer granularity and collects more detailed demographic information relative to the Current Population Survey, it enables researchers to document a much more nuanced picture of migration flows. However, the American Community Survey began collecting complete information in 2005; therefore, the time span of this survey is much shorter than the time span of the Current Population Survey. Also, the reasons for moving and the distance moved are not available in the American Community Survey.

The Internal Revenue Service migration data are a source of area-to-area migration flows at the county and state levels. The Internal Revenue Service collects annual migration information based on records of all individual income tax forms filed each year, which makes the Internal Revenue Service migration data unique and more complete as compared to other data sources. Movers in the Internal Revenue Service data are defined as tax filers who do not maintain the same address for two consecutive years. The Internal Revenue Service migration data covers about 87 percent of U.S. households (Molloy et al., 2011). However, since the data represents people who file taxes in consecutive years, the data likely does not include poor, wealthy, and elderly individuals who are not required to file taxes (Gross, 2005). Also, the Internal Revenue Service migration data does not include detailed demographic characteristics on migration flows except for income and age.

Scholars have noted that all these migration data share an important limitation (Kaplan & Schulhofer-Wohl, 2017; Molloy et al., 2011). Similar to any survey-based data, the datasets mentioned above are released with substantial lags, which makes it difficult to adjust policies for current migration trends in a timely fashion. As a result, policy analysts are increasingly interested in nontraditional real-time data that could enhance the analysis of migration trends and help to increase the accuracy of short-term migration flows predictions.

1.2 Background of Online Search

The U.S. Internet access rate has been consistently increasing, and the web today is a default channel for information search. Figure 1 illustrates the growth in percentages of U.S. adults who use the Internet over time. According to the Pew Research Center (2021), more than 9 in 10 American adults used the Internet in 2020, a strong increase from 52 percent of Americans who used the Internet in 2000. In February 2012, 73 percent of all Americans used search engines, an increase from 52 percent in January 2002 (Purcell et al., 2012). In 2012, 59 percent of American Internet users used a search engine, an increase from 30 percent in 2004. The U.S Census Bureau (2021) also provides statistics on Internet access rates by state (see Figure 2). In 2018, the highest Internet access rate was 90% for Utah, and the lowest Internet access rate was 76% for Mississippi. Even though disparities in Internet access exist among states, web access has grown dramatically across the United States over the past two decades.

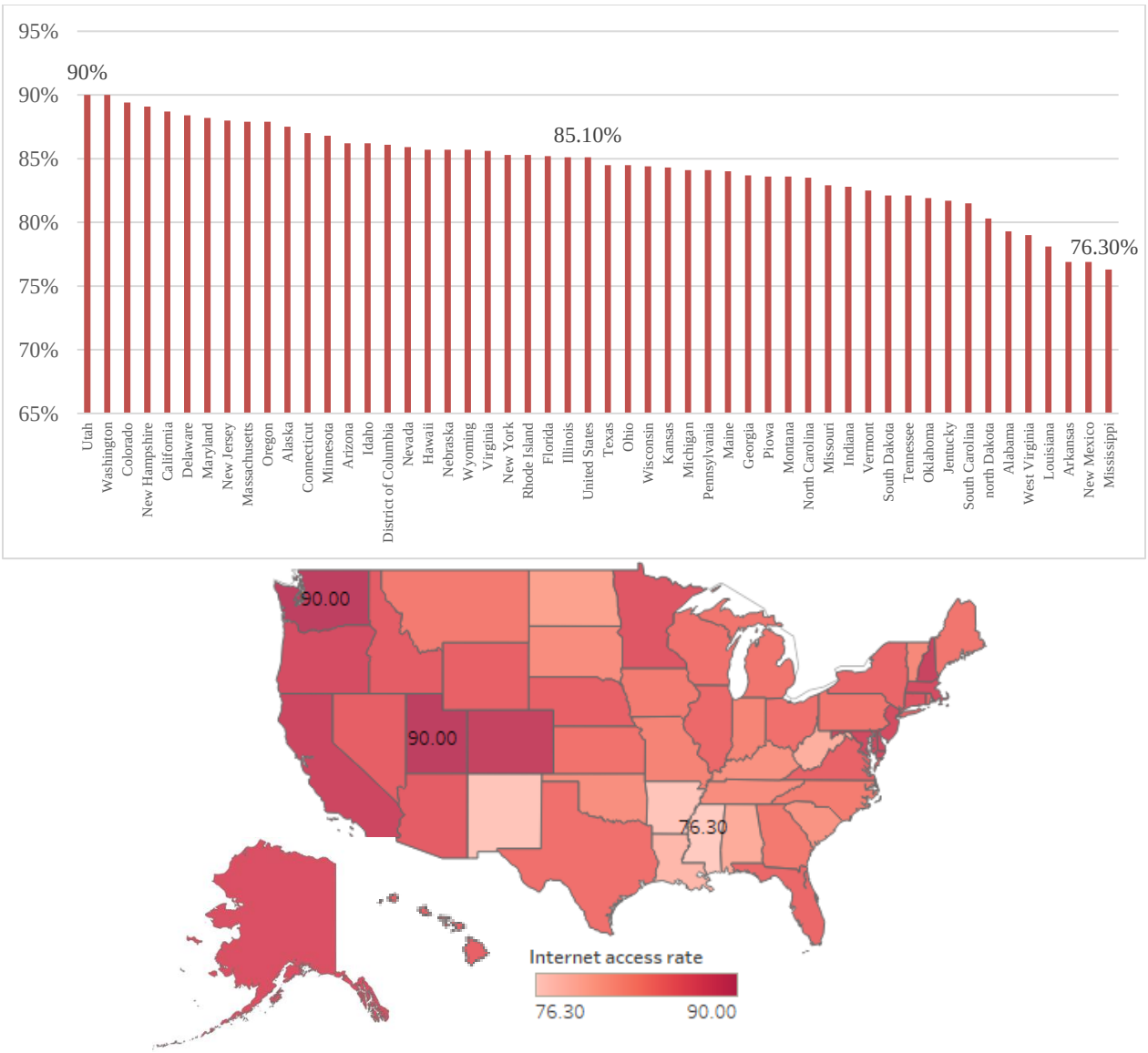


Source: Pew Research Center, Surveys of U.S. adults

Figure 1: Internet Access Rates in the United States

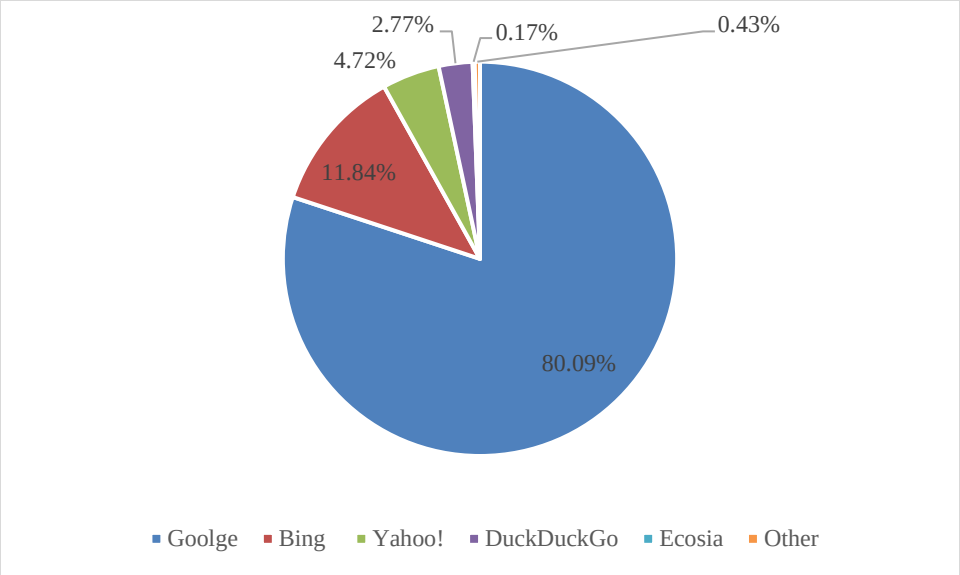
To use a search engine, users enter keywords into the search engine, and it retrieves relevant web pages and information. Users typically do not pay for this service because search providers' main revenue comes from advertising. As the most popular search engine, Google has widened the usage gap between itself and other search engines. Purcell et al. (2012) found that 83 percent of search engine users used Google in 2012, considerably more often than Yahoo! (6 percent). These statistics represent a strong increase from 47 percent of users relying on Google versus 26 percent of users relying on Yahoo! in 2004. According to Statcounter, as of August 2021, Google's desktop search engine market share in the United States is a dominant 80.09 percent, followed by Bing (11.84 percent), Yahoo! (4.72 percent), and DuckDuckGo (2.77 percent) (Figure 3).[1] Google's dominance is much greater in terms of the mobile search engine market share in the United States, where it has increased to 94.24 percent.

[1] https://gs.statcounter.com/search-engine-market-share/all/united-states-of-america



Source: U.S. Census Bureau, American Community Survey

Figure 2: Internet Access Rates in the United States by State (2018)



Source: Statcounter

Figure 3: Desktop Search Engine Market Share in the United States

The Internet and search engine use have become a staple of people's daily lives and, arguably, one of the most common and important activities related to information search and interpersonal communication for both work and private life. All these online activities leave a trace of data that can serve as input for statistical analyses. For example, search term volumes can be used to measure and predict consumers' purchase intentions. Past studies demonstrated that search volumes, i.e., the aggregated measures of popularity of specific terms submitted to a search engine, can be effective in gauging Internet users' interests or concerns (Varian, 2014).

The use of big data from online search engines has become increasingly common in many research areas. Such data has proven quite useful in its ability to enhance the analyses based on conventional data sources, and the new applications of online search data continue to proliferate. Online search data hold several advantages for research. Meshcheryakov (2018) summarized five advantages of Google Trends (i.e., Google's measures of search trends) data: (1) it is comprehensive, (2) it is quick and easy to access, (3) it provides more insightful and

accurate measures of directly unobservable variables, (4) it offers real-time and easy-to-customize data at nearly zero cost, and (5) it is offered in precompiled and user-friendly format. In addition, Google Trends provides geolocalized (or georeferenced) data. Google stores the date when the search term was entered and the location of the search users' computer.[2] With the date and the location, Google Trends illustrates trends associated with search terms in specific geographic areas over a specific period. This unique data feature allows me to construct a measure of migration intentions and analyze migration flows between any pair of the origin and destination states.

In this dissertation, I propose the use of online search data to measure migration intentions and predict interstate migration flows. I combine the IRS migration data with Google Trends, i.e., the keyword search reports that offer insights based on actual Google search statistics. These georeferenced online search data allow me to identify migration intentions of people by the state of residence, which I assume indicates the state from which the move originates. Using Google Trends data, I can assess the search intensities of specific keywords that indicate moving intention in the origin areas, both including migrants' destination choices and disregarding migrants' destinations choices. I test how successfully such measures of moving intentions predict the subsequent actual aggregate migration flows from origin to destination observed in the IRS migration data. In short, this study investigates the predictive power of online search data in forecasting the U.S. migration flows between states.

[2] https://policies.google.com/technologies/location-data?hl=en-US

1.3 Objectives of the Dissertation

The objective of this dissertation is to extend the body of knowledge regarding domestic migration in the United States by (1) introducing a new measure of migration intention and (2) evaluating the predictive power of the new measure in forecasting the short-term migration flows. The specific two research questions that I investigate are:

1. Can search engine data provide information about migration intentions from an origin state?

2. Can search engine data improve the predictive power of migration flow forecast models?

While several recent studies examined the determinants of international migration, I focus on domestic migration in the United States. The United States citizens exhibit higher mobility than citizens in other developed countries, making them ideal subjects to test the predictive power of the new measures of migration intentions. To the best of my knowledge, this study is the first to use Google Trends data in the context of domestic migration.

1.4 Contribution of the Dissertation

The contribution of this paper to the existing literature on domestic migration and online information search can be summarized in two points. First, since my approach demonstrates how to use the near real-time data to predict short-term migration flows, it could benefit policymakers and migration analysts who otherwise are confined to working with obsolete data. Second, while I chose to demonstrate my proposed analytical framework in the context of domestic interstate migration, the same approach can be applied to migration in either a more narrow or a broader

geographic context. Moreover, a similar framework could be replicated in other research areas where the availability of forecasts based on the most current data is crucial to making effective decisions and/or policy.

1.5 Organization of the Dissertation

The rest of this dissertation is organized as follows. Chapter 2 provides the theoretical background and hypotheses of the dissertation by reviewing the previous literature on internal migration and online search data. Chapter 3 presents the data, methodology, and models. Chapter 4 discusses the empirical results. Finally, Chapter 5 concludes the dissertation with final remarks and implications.

CHAPTER 2

LITERATURE REVIEW

2.1 Overview

The focus of this dissertation is on illustrating how online search data can be used to measure internal migration intentions and forecast internal migration flows. Chapter 2 will provide the dissertation's theoretical framework for both migration and online search. The first section (2.2) of the literature review begins with an account of research related to migration theory. I then review studies on the determinants of migration. The second section (2.3) reviews internal migration in the United States over the past two decades and provides qualitative evidence about why people move. The third section (2.4) reviews studies on online consumers' information search behavior, focusing on search engines and social media. In the fourth section (2.5), I discuss how information technology influences the decision to migrate. Lastly, the fifth section (2.6) provides our set of hypotheses in this dissertation.

2.2 Migration Theories

Migration is a complex phenomenon influenced by social, economic, cultural, and political factors (Molloy et al., 2011). Migration directly changes population size and structure and indirectly changes housing and labor markets. Although the concept of migration is hard to explain in one paragraph, many studies have adopted the definition from Lee's (1966) seminal paper:

> Migration is defined broadly as a permanent or semipermanent change of residence. No restriction is placed upon the distance of the move or upon the voluntary or involuntary nature of the act, and no distinction is made between external and internal migration. Thus, a move across the hall from one apartment to another is counted as just as much an act of migration as a move from Bombay, India, to Cedar Rapids, Iowa, though, of course, the initiation and consequences of such moves are vastly different. However, not all kinds of spatial mobility are included in this definition. Excluded, for example, are the continual movements of nomads and migratory workers, for whom there is no long-term residence, and temporary moves like those to the mountains for the summer. (p. 49)

In Lee's definition, there are three important components of migration: change of residence, duration of stay in the destination, and the distance between origin and destination. First, migration involves a change in one's usual place of residence. Second, the duration of stay in the destination differentiates between a permanent or semipermanent move as opposed to a temporary one. Lastly, while migration requires some distance of movement, the magnitude of distance between origin and destination does not matter.

The definition of migration developed by the U.S. Census Bureau (2021, December 3) borrows heavily from Lee:

> Migration and geographic mobility both refer to the movement of people from one location to another. Migration typically refers to moves that cross a boundary, such as a county or a state line, and is either domestic migration (movement within the United States) or international migration (movement between the U.S. and other countries). Mobility includes both short and long-distance moves. (para.1)

The U.S. Census Bureau offers parallel definitions of a migrant in its surveys based on questions about living arrangements. The ACS asks, "Did this person live in this house or apartment one year ago?" whereas the CPS asks, "Was the reference person living in this house (or apartment) one year ago?" Migration typically refers to long-distance moves, while residential mobility refers to short-distance (local) moves. Distance of movement is not typically factored in; however, change of residence and a minimum duration of residence in the destination of one year are required to classify someone as a migrant.

Many studies have attempted to study migration phenomena, creating macro and micro-analytical models based on the definition of migration. Prominent frameworks include the push-pull model from population studies (Lee, 1966), the mathematical gravity model (Dorigo & Tobler, 1983), life course theory from the social sciences (Elder et al., 2003), and the neoclassical economic theory of consumer rational decision-making (Harris & Todaro, 1970). Even though the push-pull model has come under criticism for being too simplistic and determinist (de Haas, 2011, 2021; Skeldon, 1990), it provided an intuitive and empirically grounded idea for other migration theories (Van Hear et al., 2018). Since this dissertation focuses on a variety of migration factors, I use the push-pull model as a theoretical framework.

Laws of Migration

As the foundation of modern migration research, Ravenstein's "laws" of migration are vital to any discussion of Lee's push-pull model. Owing to his studies of internal migration patterns in the United Kingdom in the 1880s, Ernst Georg Ravenstein became the first scholar to describe with scientific rigor certain empirical regularities and observations regarding migration (Ravenstein, 1885, 1889). He analyzed the census data from North America, Great Britain, and

other countries in Europe to conceptualize seven laws or principles of migration. Grigg (1977) summarized, expanded, and classified Ravenstein's "laws" of migration as follows:

1. The majority of migrants go only a short distance.
2. Migration proceeds step by step.

   Ravenstein wrote in 1885

   "… the inhabitants of the country immediately surrounding a town of rapid growth flock into it; the gaps thus left in the rural population are filled up by migrants from more remote districts, until the attractive force of one of our rapidly growing cities makes its influence felt, step by step, to the most remote corner of the Kingdom,"
3. Migrants going long distances generally go by preference to one of the great centers of commerce or industry.
4. Each current of migration produces a compensating counter current.
5. The natives of towns are less migratory than those of rural areas.
6. Females are more migratory than males within the Kingdom of their birth, but males more frequently venture beyond.
7. Most migrants are adults: families rarely migrate out of their county of birth.
8. Large towns grow more by migration than by natural increase.
9. Migration increases in volume as industries and commerce develop and transport improves.
10. The major direction of migration is from the agricultural areas to the centers of industry and commerce.
11. The major causes of migration are economic. (Grigg, 1977, p. 42-43)

Ravenstein's laws imply that (1) the volume of movement between two places increases as distance decreases (law 1); (2) migration occurs in stages (stepwise migration), as from farm to village, village to town, and town to city (Conway, 1980; Nilsson, 2003; Riddell & Harvey, 1972) (law 2); (3) individual characteristics affect an individual's migration decision (law 6 and 7); (4) technological advancements influence the decision to migrate (law 9); and (5) both desirable and undesirable conditions in origin and destination influence migration.

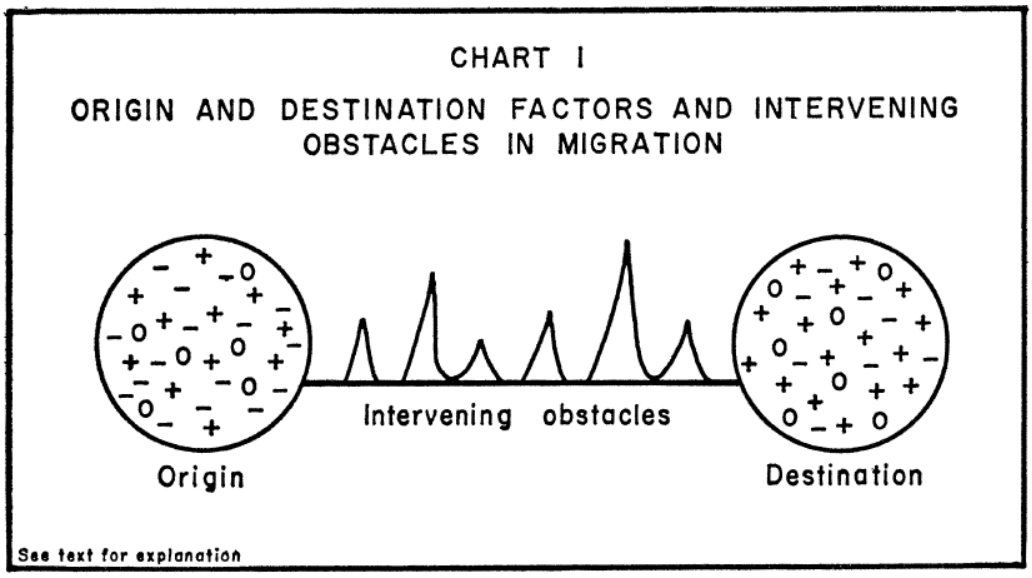
Ravenstein's generalized principles of migration inspired researchers to develop other migration theories and to examine the influencing factors of migration patterns. The eleventh law provided the basis for several economic migration theories, including the push-pull theory, and the first and third laws provided the basis for the gravity model of migration. The third, fifth, eighth, and tenth laws provided the basis for the rural-urban migration study, while the fourth law provided the basis for the return-migration study.

Push-Pull Theory of Migration

The push-pull theory was developed by Everett Lee (1966) based on Ravenstein's laws. Lee's push-pull theory suggested that there were four major groups of determinants of migration:

1. Factors associated with the area of origin
2. Factors associated with the area of destination
3. Intervening obstacles
4. Personal factors (p. 50)

He illustrated these four factors with simple marks like the plus sign, "+" (pull), the minus sign, "-" (push), and the number zero, "0" (indifference) (see Figure 4).



Source: Lee (1966)

Figure 4: Push-Pull theory

Areas of origin and destination both have "push" factors that repel people from the migration process and "pull" factors that hold or attract people to it. Push factors include decreased economic opportunities, low quality of education, low average income levels, etc., while pull factors include increased economic opportunities, high quality of education, high average income level, etc. Intervening obstacles (distance, misinformation, etc.) between origin and destination may prevent people from migrating or reduce migration flow. Individuals conduct cost-benefit analyses in which they weigh the push and pull factors against each other and additionally consider personal factors; if push factors in the area of origin or pull factors in the area of destination outperform intervening obstacles, people will migrate.

Like Ravenstein's laws, Lee's push-pull theory is one of the most popular theoretical frameworks in the study of migration. Dorigo and Tobler (1983) expressed the push-pull theory in a mathematical form. Past papers that relied on the push-pull framework examined the influence of factors such as (1) housing (Ferreira et al., 2010; Speare, 1974; Stawarz et al., 2021),

(2) employment (DaVanzo, 1978; Molloy et al., 2017; Saks & Wozniak, 2011), (3) demographics, e.g., age and gender (Champion & Fotheringam, 1998); (4) socioeconomics, e.g., education (DaVanzo, 1983; Piras, 2021), marital status (Graves & Linneman, 1979), and family ties (Mincer, 1978); (5) environmental factors, e.g., amenities (Roback, 1988; Graves, 1983; Winkler & Rouleau, 2020), and climate (Andrienko & Guriev, 2004; Roback, 1988; Winkler & Rouleau, 2020), and (6) political preferences (Mayda et al., 2018). Many of these studies attempted to identify the determinants of migration within the push-pull framework, a comprehensive theory that postulates that each place is characterized by both positive and negative factors that either attract or repel people from the area (Skeldon, 1990).

Gravity Model of Migration

The gravity model was initially an empirical one; however, researchers have added a theoretical foundation to it (Anderson, 1979; Isard, 1960; Krugman, 1980; Tinbergen, 1962). Researchers derived the gravity model from Isaac Newton's law of gravitation, modifying it to predict bilateral trade flows in information, commodities, or migration. In the gravity model of migration (Ravenstein, 1885, 1889), two areas are attracted to each other in proportion to their population sizes and the distance between them, as planets are drawn to each other in proportion to their mass and proximity. The gravity model of migration is expressed as

$$I_{ij} = k^* (P_i P_j) / D^b_{ij},$$

where I_{ij} is the interaction between places *i* and *j*, *k* is a constant, *Pi* and *Pj* are measures of the size of places *i* and *j* (e.g., populations, unemployment rates, degree of urbanization, amenities, and public expenditure), D^b_{ij} is the distance between places *i* and *j*, and *b* is the friction of

distance. According to the gravity model, the number of migrants between two places *i* and *j* decreases when the distance between them increases or the size of the places decreases.

The gravity model is the most common empirical formula for migration because it can accommodate multiple sets of push and pull factors (Dorigo & Tobler, 1983; Lee, 1966). In bilateral migration flows, the gravity model offers high explanatory power and ease of application due to its reliance on aggregate data (as opposed to microdata on individuals), and therefore becomes a workhorse in migration research (Anderson, 2011; Beine et al., 2016; Ortega & Peri, 2013).

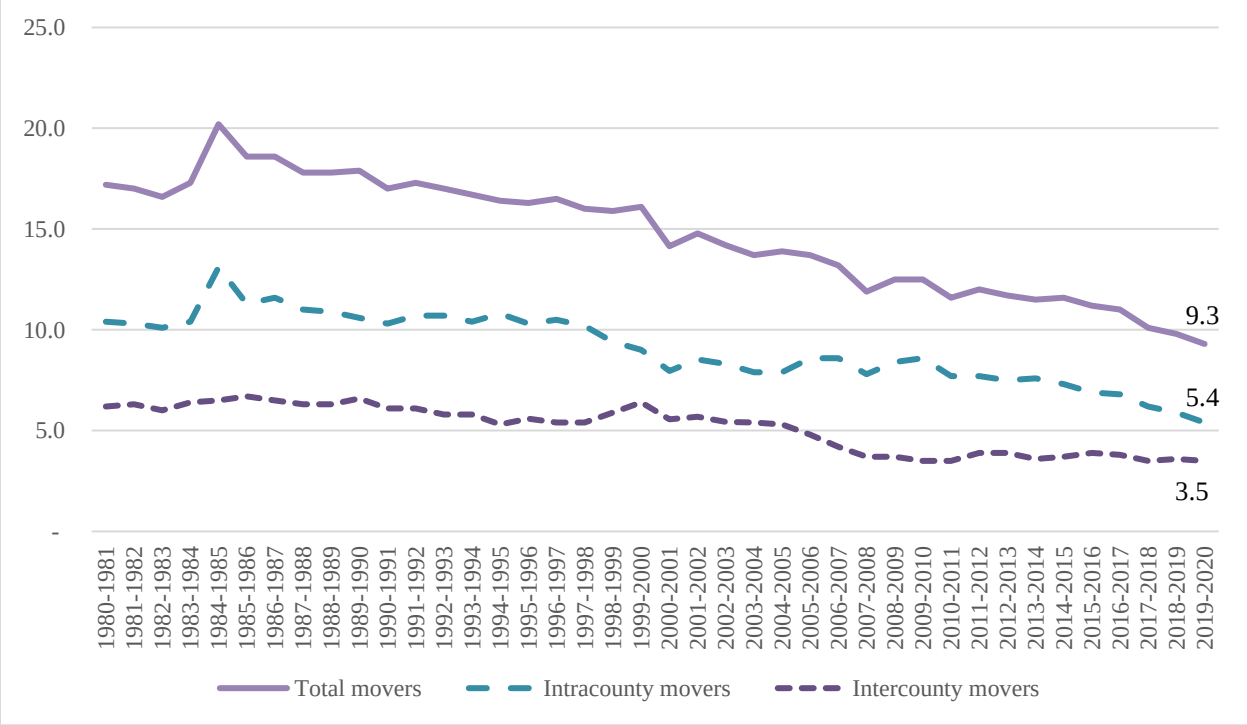
On the one hand, the gravity framework is commonly used to model bilateral migration flows with aggregated "mass" variables on origins and destinations, such as population size or GDP. On the other hand, such data used with the gravity model present difficulties in interpreting the movement of individual population members. To overcome this difficulty, Allen (1972) used the sum of individual behaviors in the mathematical form based on a random utility maximization model and applied the gravity model at the micro-level.

In this dissertation, the gravity model is used together with the push-pull theory to identify the determinants of migration.

2.3 Internal Migration in the United States

Figure 5 illustrates the evolution of U.S. internal migration rates over the past four decades. The U.S. internal migration rate, defined as the percentage of people who move in a year, has been decreasing steadily since 1980. The migration rate was 9.3 percent as of 2020, the lowest rate ever recorded. Migration changes the socio-demographic makeup of local populations, which in turn could affect both the economic and social living conditions. Declining

migration rates present state governments with significant challenges as they may cause economic downturns that adversely affect states' capacity to enact and execute public policy (Cooke, 2013; Kaplan & Schulhofer-Wohl, 2017). For example, interstate migration is an important determinant of labor and housing market trends. Frost (2020) documented that 71 percent of the increase in the number of households in Arizona in 2018 was due to domestic migration. High migration rates add flexibility to labor markets and alleviate income inequalities between areas. Moreover, migration correlates with demand for housing, which in turn affects housing and rental prices (Lin et al., 2018; Stawarz et al., 2021). Molloy et al. (2011) argued that the decline in migration rates in the 2000s resulted from the aging of the U.S. population, increased homeownership rates, decreased labor demands in some states, and economic downturns.



Source: U.S. Census Bureau, Current Population Survey

Figure 5: Annual Migration Rates in the United States

People tend to be attracted to locations with high wages or strong wage growth (Boustan et al., 2010), high housing satisfaction (Speare, 1974), or availability of affordable housing (Stawarz et al., 2021). In the CPS, survey respondents who moved in the last year are asked why they were moving. The responses are categorized into "family-related reasons" (e.g., change in marital status, a move to establish own household, a move for other family reasons), "employment-related reasons" (e.g., a move to start a new job, looking for work or lost job, a move to be closer to work or to have an easier commute, a move due to retired, a move due to other job-related reasons), "housing-related reasons" (e.g., a move motivated by desires to own a home, upgrading to a new or better house or apartment, relocating to a better neighborhood (e.g., with less crime), cheaper housing, a move due to a foreclosure or eviction, a move due to other housing reasons), and "other reasons" (e.g., cohabitation with an unmarried partner, a move to attend graduate from college, change of climate, health reasons, natural disaster, etc.).

As shown in Table 1, between 2000 and 2020, the most common reasons to move based on responses to the CPS are housing-related, followed by family- and employment-related factors. Figure 6 illustrates the percentage of people who moved for these four reasons between 2000 and 2020. The rankings of these four groups did not change over 20 years, and the percentage distribution of movers for each year did not change much either.

Table 1: Reasons for Move: 2000 to 2020

Mobility period	Family-related	Employment-related	Housing-related	Other
1999-2000	26.62	17.21	49.69	6.48
2000-2001	27.20	17.18	47.92	7.70
2001-2002	25.75	18.00	49.54	6.71
2002-2003	26.31	15.58	51.32	6.79
2003-2004	24.30	16.99	52.77	5.95
2004-2005	27.15	17.64	47.16	8.04
2005-2006	27.68	18.39	46.16	7.77
2006-2007	30.14	20.84	41.97	7.05
2007-2008	30.54	20.91	40.09	8.47
2008-2009	26.34	17.93	45.88	9.84
2009-2010	30.25	16.42	43.73	9.60
2010-2011	27.85	18.33	45.04	8.78
2011-2012	29.30	19.34	49.44	1.91
2012-2013	30.27	19.43	47.96	2.35
2013-2014	29.44	20.66	47.92	1.98
2014-2015	31.08	20.57	46.11	2.24
2015-2016	27.45	20.17	42.19	10.20
2016-2017	27.90	18.47	43.02	10.61
2017-2018	28.10	19.68	41.47	10.74
2018-2019	26.76	21.24	40.41	11.59
2019-2020	25.50	19.80	40.15	14.55

Source: U.S. Census Bureau, Current Population Survey

Employment-related reasons tend to be more important determinants for long-distance ones, while housing-related reasons explain the majority of short-distance moves (Niedomysl, 2011). The U.S. Census Bureau provides information about two different types of moves along with reasons for moving: intracounty moves and intercounty moves. Intracounty moves refer to moves within a county, while intercounty moves refer to moves across county boundaries (Schachter, 2004). Table 2 shows the percent distribution by reasons for intracounty movers and intercounty movers between 1999 and 2000[3] in panel (a) and 2019 and 2020 in panel (b). The

[3] Data source for 1999-2000 currently is not available from the U.S. Census Bureau. The numbers for 1999-2000 come from Schachter (2001).

percentage distribution of reasons does not change much between the two periods. The percentage of family-related reasons is very similar for intra- and intercounty moves in both periods. However, there is a noticeable difference between intra- and intercounty movers for employment-related reasons and housing-related reasons.

Employment-related reasons accounted for only 5.6 percent of intracounty moves between 1999 and 2000, while the same reasons accounted for 31.1 percent of intercounty moves during the same period. Furthermore, employment-related reasons for intracounty moves comprised 10.6 percent of the reasons behind moving between 2019 and 2020, while employment-related reasons for intercounty moves made up 32.4 percent during the same period. Moreover, the majority of long-distance moves, both in 2000 and 2020, were to find new jobs or to transfer jobs.

Source: U.S. Census Bureau, Current Population Survey

Figure 6: Reasons for Move: 2000 to 2020

Conversely, housing-related reasons comprised 65.4 percent of intracounty moves between 1999 and 2000, while housing-related reasons comprised 31.9 percent of intercounty moves during the same period. Also, housing-related reasons for intracounty moves made up 50.1 percent of reasons behind moving between 2019 and 2020, while housing-related reasons for intercounty moves accounted for 27.6 percent of moves during the same period (see Figure 7). The majority of intracounty movers reported that they made housing moves within a county to live in better housing. As people move less, housing became a less important factor for moving for both intra- and intercounty movers.

Table 2: Reasons for Moving by Type of Move: 1999 to 2000 and 2019 to 2020
(Numbers are in thousands)

Panel (a): 1999 to 2000

Reasons for moving	Percent distribution by reason				
	Total	Intracounty	Percent	Intercounty	Percent
Total movers	100.0	24,399	100.0	17,242	100.0
Family-related reasons	**26.3**		**25.9**		**26.9**
Change in marital status	6.2		6.2		6.2
To establish own household	7.4		9.3		4.7
Other family reason	12.7		10.4		16.0
Employment-related reasons	**16.2**		**5.6**		**31.1**
New job/Job transfer	9.7		1.4		21.6
To look for work/lost job	1.3		0.5		2.4
Closer to work/easier commute	3.5		3.0		4.2
Retired	0.4		0.1		0.9
Other job-related reason	1.2		0.6		2.0
Housing-related reasons	**51.6**		**65.4**		**31.9**
Wanted to own home/not rent	11.5		14.3		7.5
New/better house/apartment	18.5		24.2		10.3
Better neighborhood/less crime	4.4		4.8		3.9
Cheaper housing	5.5		7.5		2.8
Foreclosure/eviction	-		-		-
Other housing reason	11.7		14.7		7.4
Other reasons	**6.0**		**3.0**		**10.1**
Relationship with unmarried partner	-		-		-

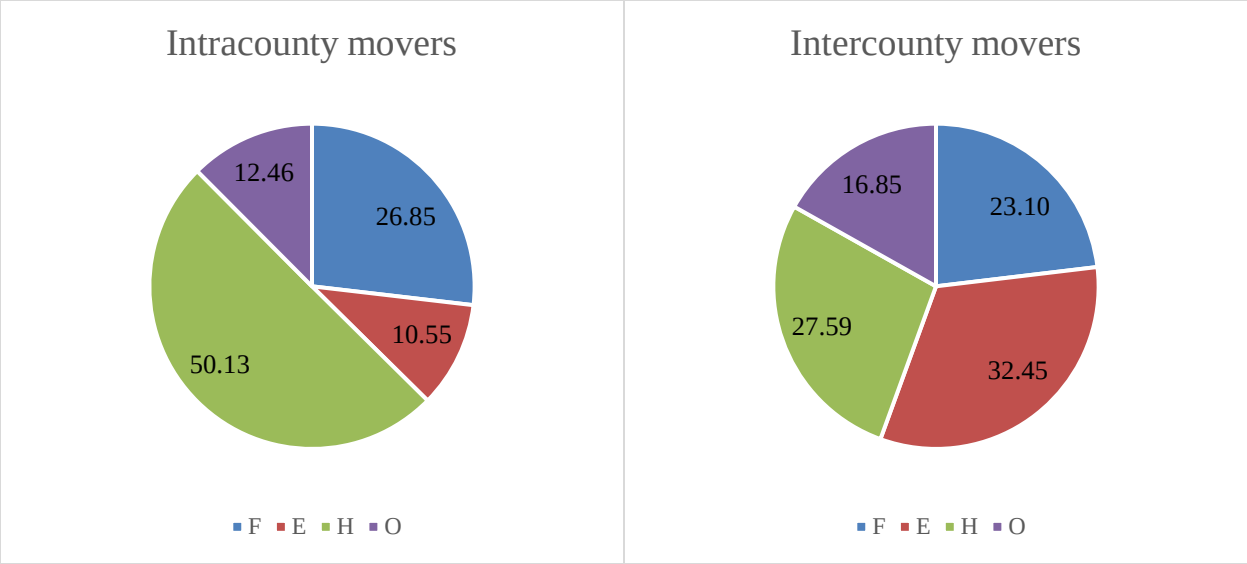
Attend/leave college	2.3	0.7	4.4
Change of climate	0.7	0.2	1.6
Health reasons	1.1	0.8	1.6
Natural disaster	-	-	-
Other reason	1.8	1.3	2.5

Panel (b): 2019 to 2020

Reasons for moving	Percent distribution by reason				
	Total	Intracounty	Percent	Intercounty	Percent
Total movers	100.0	17,522	100.0	11,292	100.0
Family-related reasons	**25.4**	**4,705**	**26.9**	**2,609**	**23.1**
Change in marital status	6.2	1,115	6.4	677	6.0
To establish own household	10.8	2,394	13.7	712	6.3
Other family reason	8.4	1,196	6.8	1,221	10.8
Employment-related reasons	**19.1**	**1,849**	**10.6**	**3,664**	**32.4**
New job/Job transfer	10.8	673	3.8	2,425	21.5
To look for work/lost job	1.6	159	0.9	308	2.7
Closer to work/easier commute	5.0	814	4.6	638	5.7
Retired	1.3	163	0.9	220	1.9
Other job-related reason	0.4	39	0.2	73	0.6
Housing-related reasons	**41.3**	**8,784**	**50.1**	**3,116**	**27.6**
Wanted to own home/not rent	8.0	1,658	9.5	659	5.8
New/better house/apartment	15.0	3,537	20.2	788	7.0
Better neighborhood/less crime	4.2	863	4.9	353	3.1
Cheaper housing	6.8	1,279	7.3	689	6.1
Foreclosure/eviction	0.7	139	0.8	51	0.5
Other housing reason	6.5	1,308	7.5	576	5.1
Other reasons	**14.2**	**2,184**	**12.5**	**1,903**	**16.9**
Relationship with unmarried partner	5.2	946	5.4	565	5.0
Attend/leave college	2.8	319	1.8	497	4.4
Change of climate	0.4	1	0.0	113	1.0
Health reasons	2.0	246	1.4	329	2.9
Natural disaster	0.4	98	0.6	13	0.1
Other reason	3.3	575	3.3	385	3.4

Source: U.S. Census Bureau, Current Population Survey[4]

[4] https://www.census.gov/data/tables/time-series/demo/geographic-mobility/p23-204.html
https://www.census.gov/data/tables/2020/demo/geographic-mobility/cps-2020.html



Source: U.S. Census Bureau, Current Population Survey

Figure 7: Reasons for Moving by Type of Movers: 2019 to 2020

As shown in Table 2, most moves are made up of local moves within a county or within a state. It is important to note that in spite of this, local moves do not significantly affect regional economic growth because there is no population growth. However, since interstate migration changes the population of states, it significantly affects housing and labor markets in each state; interstate migration is also reflected in the labor force participation rate and housing stock (Frost, 2020).

2.4 Information Search

Stigler (1961) postulated that consumers make better decisions given more information. Information search allows consumers to reduce perceived risk and uncertainty before buying a product (Dowling & Staelin, 1994). However, since information search imposes both monetary as well as non-pecuniary costs, consumers' search for information is subject to limits. Consumers will stop searching when the marginal benefits of extra information are equal to the

marginal costs of information (Guo, 2001; Hauser et al., 1993; Ratchford, 1982). The type of product or service sought affects the appraisal of costs and benefits of the information search process.

Economics and marketing sciences differentiate between three types of goods: search goods, experience goods, and credence goods (Ford et al., 1988; Nelson, 1970). Search goods are products or services with characteristics easily evaluated before purchase or consumption. In contrast, experience goods are products or services with characteristics that can be evaluated only after purchase or consumption. Finally, credence goods are products or services with characteristics that make them difficult or impossible to evaluate even after purchase or consumption. Nelson (1970) documented that information regarding experience goods (e.g., theme parks, holidays, travel, etc.) is less likely to be searched by consumers compared to information necessary to evaluate search goods (e.g., sporting goods, cameras, furniture, etc.) because the characteristics of such goods make the return on information search less valuable.

Online Search

The Internet allows consumers to pay almost nothing for information to compare products or services; as a result, the Internet has become a default channel for information search (Cole et al., 2003). However, it may not always be true that the Internet provides more information for better decision-making due to "information overload" (Lee & Lee, 2004). Nevertheless, search engines significantly reduce search costs by retrieving a vast number of web pages, displaying personalized information stored on the Internet, and enabling consumers to access information immediately at zero cost (Kumar et al., 2005). Jansen et al. (2008) stated that 80.6% of search keywords on search engines have informational intent, compared to 10.2% with

navigational intent and 9.2% with transactional intent.[5] Chen et al. (2014) documented that search engines reduce search time in comparison to offline searches.

Accordingly, the search engine market for retrieving information is growing rapidly. As Stigler (1961) stated, the Internet (or a search engine) increases access to high-quality information for consumers and thus leads to better decision making (Peterson & Merino, 2003; Ratchford et al., 2003). To make use of online information, consumers must learn how to obtain and consider their choice of search engines and strategic keywords (Kumar et al., 2005). The areas of scholarly research that have conducted studies on online searches and search engines are computer sciences, economics, consumer behavior, and marketing. Many studies have explored the choice, performance, and effectiveness of various search engines by examining their strengths and weaknesses (Bradlow & Schmittlein, 2000; Ding & Marchionini, 1996; Dong & Su, 1997; Jansen & Molina, 2006; Jansen et al., 2005; Kumar & Lang, 2007; Su, 1998).

Big Data

The increases in Internet use, especially the widespread popularity of search engines and social network platforms, have prompted researchers to use big data to forecast a variety of outcomes. Although the precise definition is difficult to conceptualize, big data comprise datasets containing larger volumes, greater variety, and more velocity (the "three Vs") relative to conventional data. Traditional techniques cannot manage these datasets effectively (Constantiou & Kallinikos, 2015; Elgendy & Elragal, 2016; Mavragani et al., 2018). Oracle Corporation, the

[5] Definitions of classifications of Web queries: (1) informational: queries meant to obtain data or information in order to address an information need, desire, or curiosity; (2) navigational: queries looking for a specific URL; (3) Transactional: queries looking for resources that require another step to be useful

second-largest software company in the world, defines traditional data as "structured and stored in a relational database," which can be managed from one computer, while defines big data as " unstructured and semistructured data types, such as text, audio, and video, required additional preprocessing to derive meaning and support metadata." (Oracle, para.2)

Information derived from Internet search queries and activities on social networks can help consumers make better decisions and allow analysts and decision-makers to generate insights into consumer needs and desires and predict their future behavior (Elgendy & Elragal, 2016). For example, Ettredge et al. (2005) used search engine data to predict U.S. employment rates. They assumed that "people reveal useful information about their needs, wants, interests, and concerns via their Internet behavior, and that terms submitted to search engines reflect this information" (p. 87). The authors used six job-related search terms — "job search", "jobs", "monster.com", "resume", "employment", and "job listings" — and found that search volumes significantly correlate with official unemployment statistics.

Two seminal papers demonstrated the applicability of search engine data to describe search users' intentions and predict relevant outcomes. Choi and Varian (2009a, b) used the Search Volume Index (SVI) of search terms provided by Google Trends to predict retail sales, home sales, automotive sales, tourism trends, and initial claims for unemployment benefits. They provided evidence that models with the SVI had greater predictive power relative to models without this covariate. Ginsberg et al. (2009) used 45 search terms from an early version of Google Flu Trends to estimate the regional and state incidence of influenza infections with a lag of one day. These two papers served as inspiration for researchers to conduct similar analyses in areas such as house prices (Beracha & Wintoki, 2013; Wu & Brynjolfsson, 2015), automobile sales (Kuruzovich et al., 2008), stock prices and trading volumes (Da, Engelberg, & Gao, 2015;

Joseph et al., 2011; Preis et al., 2013), unemployment rates (Askitas & Zimmermann, 2009; D'Amuri & Marcucci, 2017), real estate investment activities (Gupta & Das, 2022), and disease outbreaks (Carneiro & Mylonakis, 2009).

The advent of social media networks, such as Facebook and Twitter, has opened new avenues for the research community to analyze data generated by users of such services (Schoen et al., 2013). Carr and Hayes (2015) define social media as "Internet-based, disentrained, and persistent channels of mass personal communication facilitating perceptions of interactions among users, deriving value primarily from user-generated content" (p. 49). Rousidis et al. (2019) reviewed the literature that relied on data collected from social media, primarily Facebook, Twitter, YouTube, and LinkedIn, and documented that such data were most commonly used to predict stock market movements (Pagolu et al., 2016), real estate prices (Zamani & Swhartz, 2017), other product prices and/or promotions (Hudson et al., 2016), movie hits (Asur & Huberman, 2010), election outcomes (Gayo-Avello, 2012; Isotalo et al., 2016), disease outbreaks (Subramani et al., 2018), weather (Rossi et al., 2018), and damage from natural disasters (Chen et al., 2017; Sadilek et al., 2016).

Because social media users can create and share content interactively (Schoen et al., 2013), social media data are useful in conducting sentiment analysis with a prediction model (Iftikhar & Khan, 2020). Recent studies have focused on prediction modeling to reduce theoretical and methodological issues arising in social media data. The most well-known challenge of social media data is self-selection bias, in which social media users do not constitute a representative sample of the population (Schoen et al., 2013). Self-selection bias leads to biased estimates and thus results that cannot be generalized (Winship & Mare, 1992). To adjust for self-selection bias, predictive models can incorporate the propensity matching technique

(Schonlau et al., 2009), sample weighting adjustment (Bethlehem & Biffignandi, 2012), or adopt a controlled experimental (Kohavi et al., 2009) or quasi-experimental (Oktay et al., 2010) designs.

A unique feature of big data derived from search engines and social media is its applicability to the so-called "nowcasting," namely the prediction of current or imminent future outcomes (Banbura et al., 2010). According to Schoen et al. (2013), traditional forecasting models provide lead-time estimates, whereas nowcasting models provide real-time estimates. The lead-time prediction usually requires the researcher to specify the future time horizon of the forecast, and the further the forecast reaches into the future, the more valuable the prediction. The precise definition of a prediction horizon may not be needed for nowcasting, as this type of prediction is intended to offer insight into phenomena or behaviors that are already taking place but are perhaps not yet fully observable or measurable. For example, individuals who are in the process of making the moving decision engage in information search behaviors and make arrangements that culminate in the move, but the actual relocation is not observed until after the fact. Many researchers have turned to nowcasting models using online search and social media data to avoid the release lag associated with other secondary data sources (Carrière-Swallow & Labbé, 2013; Choi & Varian, 2012; Lampos & Cristianini, 2012). In this dissertation, I aim to demonstrate the applicability of nowcasting to the analysis of migration flows.

2.5 Migration and Online Searches

The Internet might play an important role in the decision-making process for moving. Historically, migrants had to visit a potential destination or consult with relatives or friends to obtain information about their potential relocation target (Banerjee, 1984). The Internet,

however, has significantly reduced information search costs (Vilhelmson et al., 2013). Likewise, the Internet eliminated or reduced the need for travel and diminished the psychic costs related to the uncertainty of living conditions (Choo & Mokhtarian, 2007; Kotorri et al., 2020; Winkler, 2017). Furthermore, the Internet weakens the importance of push factors, reducing the incentive for migration. For example, an increase in foreign direct investment flows through the Internet[6] (Choi, 2003; Yin & Choi, 2021) and the demand for workers with internet-related skills[7] (Akerman et al., 2015) leads to economic growth and creates jobs in the origin area (Choe, 2003; Czernich et al., 2011). In contrast, the Internet weakens pull factors by enabling workers to work remotely (Agrawal et al., 2015).

Data from online searches appear to be well-suited to predict migration patterns. Since migration occurs between origin and destination, georeferenced data are essential for migration flow analyses (State et al., 2013). Zagheni and Weber (2012) introduced the innovative approach of using georeferenced data from several social networks to estimate international migration rates. They analyzed about 34 million Yahoo! e-mail messages by IP address to identify each user's geographic location and found that estimated migration rates were consistent with official migration statistics from the respective countries. In a similar study, State et al. (2013) used georeferenced data from more than 100 million Yahoo! e-mail messages to identify tourists and migrants and to estimate migration probability between countries. They defined migrants as people who spent at least 90 days in a foreign country, while tourists were defined as people who visited foreign countries for less than 90 days. These studies presented evidence that

[6] Yin & Choi (2021) argue that the Internet can attract foreign direct investment because "the Internet reduces the transaction and production costs for foreign investors, and provides them with more access to information about alternative investment opportunities."
[7] Akerman, Gaarder, & Mogstad, (2015) found that a substantial increase in the adoption of IT-enhanced technology in firms increases the demand for technical and problem-solving.

georeferenced data can provide useful information to estimate migration patterns, especially in situations where other available data suffer from defects such as release time lag or inconsistent definitions/measures of migration.

Zagheni et al. (2014) used georeferenced Twitter data from mobile GPS devices or computer IP addresses for approximately 500,000 users in OECD countries to estimate international and internal migration flows and found their estimates consistent with official migration statistics. They concluded that georeferenced data available ahead of official migration statistics could inform predictions of short-term migration. The same method was applied by Fiorio et al. (2017) to estimate short-term and long-term migration flows in the United States.

While informative, the georeferenced social network data is not free of bias (Hecht & Stephens, 2014; Malik et al., 2015; Zagheni & Weber, 2015). Given that young people use social network platforms more frequently than older people, the data may suffer from selection bias and may not be fully representative of the general population. To mitigate this selection bias, State et al. (2014) used the LinkedIn service to track the migration histories of professionals and therefore accounted for a broader cross-section of age distribution. LinkedIn users, however, may misrepresent the general population in terms of characteristics such as educational attainment and work experience. Zagheni et al. (2017) estimated the "stocks of migrants" based on Facebook's advertising platform, which identifies foreign-born migrants to target them with customized ads. The authors demonstrated that Facebook-generated estimates of foreign-born migrants in the United States were highly correlated with estimates from the ACS data.

The data provided by search engines might be particularly useful and well-suited to the analysis of migration trends as Internet search engine users represent a much broader cross-section of the population relative to social network users. Böhme et al. (2020) argued that the

search term volumes related to migration could provide real-time input information needed to gauge moving intentions. They used the Google Trends Search Volume Index (SVI) of 67 search terms (65 single keywords and two phrases) in three languages to estimate international migration flows among OECD countries, demonstrating that search term volumes measured in origin areas can improve the precision of estimates of international migration flows. In a similar study, also using the Google Trends data, Wanner (2021) predicted short-term migration inflows to Switzerland based on one key phrase, "working in Switzerland," in four languages (German, French, Italian, and Spanish). Lin et al. (2019) forecasted the U.S. interstate migration trends using search term data from Microsoft's Bing.com. The authors focused on housing and employment queries and showed that younger and lower-income people were more likely to migrate for employment-related reasons, while older and higher-income people were more likely to migrate for housing-related reasons. They also found that different geographic areas experienced migration for varying reasons. For example, both Florida and Texas have positive net migration inflows. However, Florida received more online search interest related to housing migration, whereas Texas had more online search queries on employment-related migration.

2.6 Hypotheses Development

Because official migration data are released with a significant lag, it is difficult for policymakers to adjust housing and labor market policies for current and future migration trends. To address this problem, forecasting and nowcasting with data from search engines and social media have attempted real-time predictions for international and internal migration. In particular, some predictive models have used Google Trends data as a proxy for migration intention, an important predictor of future migration (Böhme et al., 2020; Wanner, 2021).

Many studies documented that there is an empirical association between migration intentions and actual migration (De Jong, 2000; De Jong et al., 1985; Kley & Mulder, 2010; Kley, 2011; Lu, 1998; Van Dalen & Henkens, 2013). However, as few datasets combine migration flows with migration intentions, most studies on migration intention and behavior rely heavily on classical surveys with small sample sizes. As the literature review reveals, online search engine data offer possible solutions to the small sample problem because these datasets can measure numerous search users' interests and intentions for a variety of real-time activities.

I expect that search term volumes related to migration will inform predictions of current and short-term trends of imminent future interstate migration in the United States. This expectation leads to the following set of hypotheses:

Hypothesis 1: Google searches related to migration are associated with interstate migration flows in the United States.

Hypothesis 2: Google searches related to migration predict interstate migration flows in the United States.

CHAPTER 3

DATA AND METHODOLOGY

This dissertation investigates search term intensities related to migration in the context of internal migration. Given the aforementioned characteristics of search terms and search engine use, I argue that the measures of online search intensity can capture current trends in interstate migration flows in the United States. That is, I assume that the specific keywords used in the search queries indicate search users' interest in or concern about a destination state, and the overall volume of the terms is correlated with the observed realizations of relocation plans. Therefore, I propose the use of this measure to predict migration flows.

3.1 Data

Migration and State Data

The empirical analysis uses annual IRS migration data for the period spanning between 2004 and 2018. There are three reasons to choose IRS migration data from among the three sources of government migration datasets mentioned above. First, the IRS data cover about 87 percent of U.S. households (Molloy et al., 2011), while ACS and CPS data comprise less representative samples. In particular, ACS and CPS survey-based methods generate many data points measured as zero migration flows due to larger margins of error for smaller regions. Second, the IRS has published annual state-to-state migration flows since 1990, whereas the ACS has published annual estimates of migration rates since 2006. Google Trends has published data on search volume for specific search terms since 2004, indicating that the IRS data overlaps

the Google Trends data for a longer timeframe. Third, the IRS data provide area-to-area inflows and outflows at both the county and state levels, enabling researchers to analyze bilateral migration flows. I use state-to-state migration outflows from all 50 states and the District of Columbia and combine them with Google Trends data.

Given that IRS migration data represent persons who file taxes in consecutive years, they likely do not include poor, wealthy, and elderly individuals who are not required to file taxes (Gross, 2005). Our specifications avoid this limitation by relying on changes in migration flows over time within bilateral corridors.

I utilize several macroeconomic variables to control for the origin and destination state-level variations in the most significant push and pull determinants of migration flows. I use median household income and population size obtained from the U.S. Census Bureau as the two main gravity forces. I also include the house price index from the Federal Housing Finance Agency (FHFA) and unemployment rates reported by the U.S. Bureau of Labor Statistics to control for the condition of local housing and labor markets. The cost of living index obtained from the Council for Community and Economic Research (C2ER) is included to capture the relative affordability. I also include a dummy variable for political party strength in the U.S. based on representation in the U.S. House of Representatives, which is coded 1 if the number of votes for Republican in the U.S. representative is greater than the number of votes for Democratic and 0 otherwise. The average number of days with temperature above 90 degrees, obtained from the Parameter-elevation Regressions on Independent Slopes Model (PRISM) climate data, is used to capture environmental and weather-related differences between states. Since I assume that people use the Internet to search for moving-related information, I also

consider the U.S. Internet access rates across states obtained from the National Telecommunications and Information Administration.

Google Trends Data

Google Trends provides a measure of the relative frequency of actual search requests used with the Google search engine. Google Trends normalizes a representative sample of total search terms by time and location to create a single continuous index variable (SVI). Although Google Trends uses a sample of actual searches, the sample is sufficiently large to represent the population of the billions of search queries submitted daily. The procedure requires that each search term volume is divided by all keyword searches for a specific period within a specific geographic area. The resulting time-series data is then scaled between 0 and 100, with 100 signifying the maximum search interest for the period and geographic area selected and 0 signifying searches made by very few people. Thus, the SVI measures relative search interest for a specific keyword or phrase within a given period and geographic area. In other words, although two regions may have the same index score, their total individual search volumes may not be the same. To avoid counting redundant searches that would artificially inflate the index value, SVI does not take into account repeated inquiries from the same person conducted in a short period. Google classifies keywords into predefined categories, e.g., “Real Estate,” “Jobs & Education,” “Science,” “Sports,” etc.

Previous empirical studies on migration that utilized online search keywords used a single term (Wanner, 2021), several terms/phrases (Böhme et al., 2020), or predefined categories (Wu & Brynjolfsson, 2015). Google allows researchers to obtain data on an almost infinite

combination of online search queries unless a given keyword cannot generate search volume data due to an insufficient number of searches.[8]

Following Böhme et al. (2020), rather than selecting one term, I choose a variety of keywords related to the migration process to ascertain the search users' intent. To do so, many studies have used Wikipedia for keyword expansion to analyze semantic networks (Keikha et al., 2018; Vidal et al., 2012; Wu et al., 2017). Zhao et al. (2012) used Wikipedia to extract keywords and tested the performance of information retrieval on Google. They provided evidence that the Wikipedia-based keyword expansion approach can improve the quality of information retrieval.

I determine keywords in a Wikipedia article that can express migration intention based on push-pull theory and the migration online search literature (see Figure 8). Table 3 provides the list of main keywords. The first list of keywords includes "housing," "employment," and "climate" for unilateral migration analysis. To expand this list, I use Wikipedia's "See also" and "What links here" features, which provide semantically similar terms within Wikipedia (Zhang, 2006; Farhoodi et al., 2009; Bawakid & Oussalah, 2010; Schwarzer et al., 2016; Beringer et al., 2019). Next, I download predefined categories related to migration, such as employment and housing. To approximate overall housing and employment interest, I use the following categories: "Real Estate", "Real Estate Agencies," "Real Estate Listings," "Apartments & Residential Rentals," "Jobs," "Job Listings," "Developer Jobs," and "Jobs & Education."

I create two different types of SVIs for unilateral (from one origin state to all other destination states) and bilateral (from one origin state to one destination state) flows to test my hypotheses. Keywords for unilateral analysis are designed to measure broad-based migration

[8] Google systematically limits maximum daily download quantities per IP address.

intentions, while keywords for bilateral analysis are designed to measure specific migration intentions to a destination state.

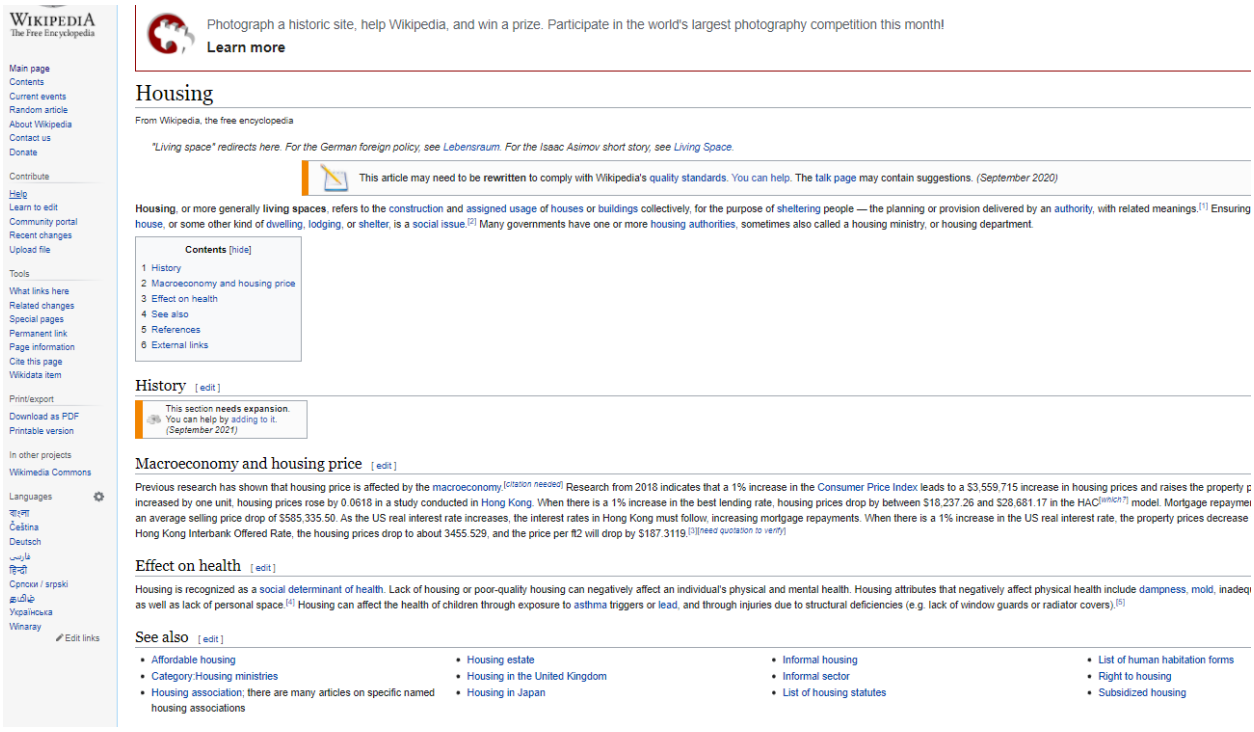
First, I obtain unilateral time series for each of my main keywords and predefined categories in a given origin state and year. 53 keywords and 8 predefined categories, over 15 years in 51 origin states (including DC), are used to create origin-specific SVI variables for unilateral analysis. Second, for bilateral migration analysis, I add the names of destination states to these 53 keywords to identify the interest expressed by a specific destination state (e.g., "house Georgia" and "job California," etc.). However, 8 predefined categories are not used for bilateral analysis because the names of destination states cannot be combined with these predefined categories. In addition, I use keywords for destination states to observe general interest in potential destinations (e.g., "Georgia"). As a result, this process creates origin-destination-specific SVI variables, which have 38,250 (51 * 50* 15) bilateral migration corridors, over 15 years in 51 origin and destination states.

People may search the above keywords for reasons other than migration interest (Böhme et al., 2020; Ormerod et al., 2014; Scharkow & Vogelgesang, 2011). However, these reasons are less problematic when a large number of search terms comprise the general intention of the topic (Böhme et al., 2020). Also, variation in search intensities over time indicates people's future interests (Scharkow & Vogelgesang, 2011).

Table 3: List of Keywords

Housing	**Employment**	**Environment**	**Predefined category***
housing	employment	climate	Real Estate
house	job	weather	Real Estate Agencies
apartment	work	temperature	Real Estate Listings
real estate	occupation	cost of living	Apartments & Residential Rentals
property	payroll	tax	Jobs
relocation	minimum wage	state income tax	Job Listings
moving	unemployment	crime	Developer Jobs
buy house	internship	school	Jobs & Education
rent house	career	college	
rent apartment	labor	university	
Mortgage	layoff	health	
Zillow	employer		
redfin	hiring		
Trulia	income		
realtor	salary		
Remax	recruitment		
neighborhood	welfare		
metropolitan	rush hour		
city	traffic		
home	commute		
foreclosure	retired**		

Notes: *Predefined category keywords are only used for unilateral migration analysis. **PageRank is an algorithm used by Google Search to rank web pages in their search engine results. Google ranks different pages for various forms of keywords, such as retire, retired, retirement, and retiring. Google PageRank even treats plural keywords differently from singular. Therefore, Böhme et al. (2020) used the Boolean operator "OR" to combine the different versions of the same keyword (e.g., applicant OR applicants). However, this method can cause reduced frequency and volume of searches when creating a search volume index. Therefore, I do not use the Boolean operator in this dissertation.



Notes: This is the Wikipedia article on "housing," which is used to expand keywords related to housing.

Figure 8: Keyword Selection from Wikipedia

Descriptive Statistics

Table 4 provides the descriptive statistics of bilateral migration panel data. The sample comprises 38,250 observations of paired origin-destination states between 2004 and 2018. On average, the bilateral migration flow comprises 1,262 migrants. The minimum number of migrants in a single bilateral flow was three[9] (South Dakota to Rhode Island in 2005 and 2006, Wyoming to Rhode Island in 2008, District of Columbia to North Dakota in 2009, North Dakota to Rhode Island in 2009, and Wyoming to Rhode Island in 2013). The maximum bilateral flow included 44,629 migrants (New York to Florida in 2017). The average of the SVI variables for

[9] Before 2010, if the state out-migration flows were based on less than three returns, then the number was not shown to protect the confidentiality of information of individual taxpayers. After 2010, only the state out-migration flows containing ten or more returns are included. 74 out of 38250 observations are not recorded (e.g., 2011: 4, 2014: 10, 2015:37, 2016:15, 2018: 8). Although I replace them with five, the minimum value is still three.

bilateral destination states was 33.66. The minimum value was zero when people searched for "District of Columbia."[10] The maximum value was 95.83 when people in California searched for "New York" in 2004.

Table 4: Descriptive Statistics of Main Variables in Bilateral Analysis

	N	Mean	SD	Min	Max
Bilateral migration flow	38250	1262.42	2652.65	3.00	44629.00
Bilateral SVI for destination	38250	33.66	16.14	0.00	93.83
Median Household Income	38250	61627.06	9758.50	35992.00	89007.00
Population	38250	6095.49	6842.25	509.11	39461.59
Unemployment rate	38250	5.82	2.10	2.40	13.70
House price index	38250	355.07	116.34	173.68	892.45
Internet user	38250	0.71	0.08	0.45	0.87
Cost of living index	38250	104.79	16.99	83.71	187.60
Political party strength	38250	0.54	0.50	0	1
No. days over 90 degrees	38250	0.44	11.04	-42.97	49.97

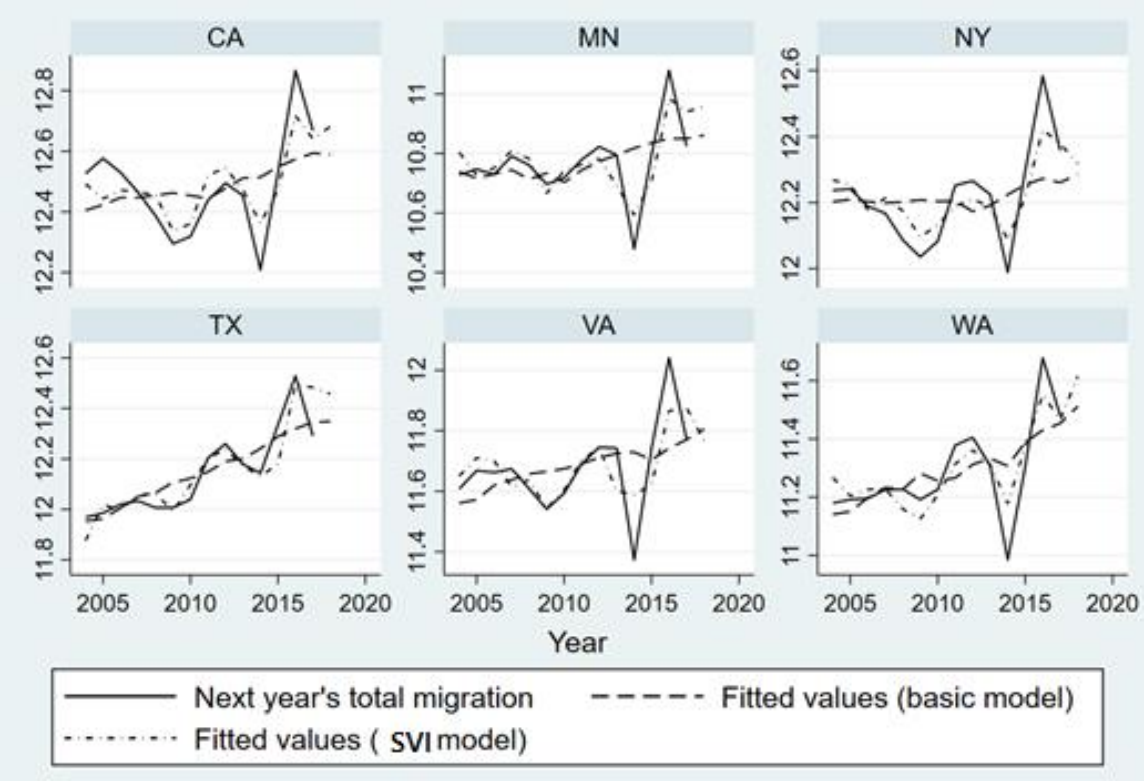
Notes: Since I use migration flows with a one-year lag in the model, 1250 observations were dropped. For the number of days over 90 degrees, I calculate the deviation from the state's 30-year average (1990-2020), subtracting the average number of days from the number of days in each year (Winkler & Rouleau, 2020).

Preliminary Check

A simple OLS regression is used to examine how closely the SVIs reflect actual migration flows in the unilateral migration specification. The dependent variable is the log of the annual unilateral migration flows (plus one). Only two gravity variables, the log of the population (origin) and the log of median household income (origin), are used in the models as explanatory variables, but the prediction also includes origin state fixed effects. Next, to examine whether the Internet search intensities improve prediction, I compare the fitted values from two OLS models: the model without the SVI variables and with the SVI variables. Figure 9 plots an arbitrary selection of six states' annual unilateral migration flows from the origin states (solid line), the fitted values from the OLS model without the SVI variables (dashed line), and the fitted

[10] As searches are made by very few people, the resulting number appears as zero. https://support.google.com/trends/answer/4365533?hl=en&ref_topic=6248052

values from the OLS model with the SVI variables (dotted line). Clearly, the fitted values from the model with the SVI variables can predict future fluctuations in annual migration flows better than those from the model without the SVI variables. In other words, population size and median household income have lower predictive power relative to the regression variate that accounts for online search interest. Overall, the plots suggest that there is a strong relationship between the SVI variables and migration flows, and the SVI variables related to migration serve as an effective indicator of migration intentions.



Notes: The graph compares three time trends, the log of annual unilateral migration flows with one-year lead and two fitted values of OLS regression with the SVI variable (SVI model) and without the SVI model (basic model).

Figure 9: Annual Migration Flows and Two Predictions

3.2 Methodology

Following Böhme et al. (2020), I test both the unilateral and bilateral model specifications with the SVI variables. These alternative model specifications are used to measure the SVI's predictive power under differing conditions. The unilateral model specification is used to provide information about the predictive power of SVI in forecasting the aggregate migration decisions from an origin state, regardless of the migrants' destination choices. Conversely, the bilateral model specification is used to provide information about the predictive power of SVI when the forecasts include information about migrants' destination choices.

Unilateral model

The following model is estimated using OLS to measure the relationship between the SVI variables and unilateral migration flows:

$$Y_{ot+1} = \beta_1 SVIuni_{ot} + \beta_2 O_{ot} + \gamma_o + \delta_t + \epsilon_{ot} \quad (1)$$

where *o* indexes the origin state and *t* indexes time. The dependent variable, Y_{ot}, is the log of annual migration flows from the origin state to all other states in the following year (plus one). By using the next year's value of the dependent variable, I address potential concerns about the endogeneity issue caused by reverse causality (Piras, 2021). This handling of the dependent variable also allows for sufficient preparation time before the move is observed in the data (Böhme et al., 2020). $SVIuni_{ot}$ denotes a vector of unilateral SVI variables. O_{ot} is a vector of time-varying origin-specific variables, γ_o denotes a vector of origin state fixed effects, and $\delta_t$denotes a vector of year fixed effects. Finally, ϵ_{ot} denotes an error term at the origin state level.

Bilateral model

The gravity model is used to estimate future bilateral migration flows. The bilateral migration flow is modeled as:

$$Y_{odt+1} = \beta_1 SVIbil_{odt} + \beta_2 SVIdest_{odt} + \beta_3 SVIuni_{ot} \times SVIdest_{odt} + \beta_4 O_{ot} + \beta_5 D_{dt} + \gamma_{ot} + \delta_{dt} + \tau_{od} + \epsilon_{ot} \quad (2)$$

where *o* and *d* index the origin and destination states, respectively. The dependent variable, Y_{odt+1}, is the log of annual migration flows from the origin state *o* to the destination states *d* observed in the following year (plus one). $SVIbil_{ot}$ denotes a vector of bilateral SVI variables, $SVIdest_{ot}$ denotes a vector of destination state SVI variables, and $SVIuni_{ot} \times SVIdest_{odt}$ is a vector of interaction terms between these variables. The interaction terms may represent different migration intentions compared to the stand-alone bilateral SVI variables and therefore they help improve the predictive power (Böhme et al., 2020). O_{ot} is a vector of time-varying origin-specific variables, D_{dt} is a vector of time-varying destination-specific variables, γ_{ot} denotes a vector of origin time fixed effects, $\delta_t$denotes a vector of destination time fixed effects, and τ_{od} denotes a vector of origin-destination pair fixed effects. ϵ_{odt} denotes an error term. A large set of fixed effects is used in gravity specifications because it eliminates the time-varying factors at the levels of origin and destination area, such as population and GDP, bilateral time-invariant factors such as distance, and unilateral migration policy changes.

Zero flows are common in bilateral data (Helpman et al., 2008), including bilateral migration data. In particular, for migration flows measured with finer granularity (e.g., city level as opposed to the national level), zeros are more likely to exist. At the same time, econometric modeling of migration flows usually requires a logarithmic transformation of data to account for adverse effects of data distribution and outliers. Since the logarithm of zero is undefined,

deleting zero observations in the log-linear OLS regression may lead to sample selection and biased coefficient estimates of the gravity model. To address this problem, several techniques are employed in empirical studies, such as (1) adding a small positive constant (Piras, 2021), (2) using limited dependent variable regressions with censored migration data (Mayda, 2010), (3) using the Poisson model, or (4) using the Heckman sample selection model.

In particular, the Poisson and Heckman models are two widely used methods to deal with the presence of zeros for estimating the gravity equation. First, Silva and Tenreyro (2006) proposed the Poisson pseudo-maximum likelihood estimator for non-linear models under weak OLS assumptions for three reasons: (1) the Poisson estimator of the gravity model is consistent in the presence of fixed effects like OLS, (2) the Poisson estimator includes zero observations, and (3) the coefficients from the Poisson model are easy to interpret. However, even though the Poisson model can produce unbiased estimates, it produces inefficient flow estimates due to over-dispersion in the dependent variable and excess zero flows (Burger et al., 2009). The second approach for dealing with the occurrence of zeros is Heckman's (1979) sample selection estimator. The sample selection bias due to zero observations can be corrected through a two-step statistical approach. The first step is to determine the probability that two areas engage in migration flows in the sample. The second step is to determine the expected value of bilateral migration flows, conditional on the existence of a migration relationship.

CHAPTER 4

EMPIRICAL RESULTS

4.1 Unilateral model

The results of estimations of parameters and standard errors from Equation 1 are reported in Table 5. Columns 1 and 2 show the results for the baseline model and the SVI model of migration outflow from the origin state to all destination states, including only two main gravity variables measured in the origin state: log population size and median household income. The coefficient on log population size is positive and statistically significant in the baseline and SVI models. I conduct a joint hypothesis test using an F-statistic to learn whether the SVI variables are jointly statistically significant in a regression predicting migration flows. The p-value of the joint hypothesis test is less than 0.001. Therefore, I can reject the null hypothesis that the vector of SVI variables has no explanatory power at any reasonable level of significance.

Since origin state and year fixed effects explain most of the variation in this model, the overall-R^2 in the baseline and SVI models is 99.5% and 99.6%, which are very high. To test the predictive power of the SVI variables, I focus on the within-R^2, which measures how much of the variation in the dependent variable within the regression observation units is captured by the model. Given that my model specifications include various configurations of observation-level fixed effects, the within-R^2 is a more appropriate metric of model fit than measures based on between-observation comparisons. The within-R^2 measures how well explanatory variables account for changes in the aggregated migration flows within each of the origin states over time. The inclusion of the SVI variables in the baseline model increases the within-R^2 from 25.2%

(column 1) to 42.1% (column 2). This implies that the SVI variables provide strong additional explanatory power in the prediction of outflow migration.

In columns 3 and 4, I estimate the analogous equations with an extended set of control variables. The coefficients on log population size, house price index, unemployment rate, and the cost of living index are positive and statistically significant. These results are consistent with previous studies that described housing and employment as the main drivers of migration (Böheim & Taylor, 2002; Henley, 1998; Jackman & Savouri, 1992; Zabel, 2012). The p-value of the test of joint significance for the SVI variables is less than 0.001. Thus, the SVI variables in the extended model are associated with unilateral migration flows.

Relative to the most parsimonious model (column 1), the inclusion of other origin-specific control variables increases the within-R^2 to 34.2% (column 3). With the inclusion of additional origin-specific control variables in the extended model (column 4), the SVI variables provide additional predictive power increasing the within-R^2 to 48%. More importantly, the SVI variables contribute sizable additional predictive power even when the extended set of control variables is included in the prediction. In sum, the SVI variables are associated with aggregate interstate migration flows and significantly improve the predictive performance of the unilateral migration model, which supports Hypotheses 1 and 2.

According to Google Search Help, search queries that do not include a keyword associated with a specific geographic location are interpreted to prioritize search results in geographical proximity to users' current location. Therefore, such location-free search terms may be more likely to capture intracounty moves or moves within the same state, suggesting that the inclusion of the same-state movers in the estimation sample could further increase the predictive power of the SVI variables. In columns 5 and 6, the dependent variable is the log of total

migration flows that include same-state moves. In columns 5 and 6, I find that both the within-R^2 and the overall-R^2 increase compared to the values in columns 1 and 2, as I expected. Also, the p-value of joint significance for SVI variables is 0.000. These results suggest that the SVI variables could provide more predictive power when search terms are more informative about targeted locations. To test this hypothesis more directly, I conduct bilateral migration estimation in the next section.

Table 5: Results of Unilateral Regression

	(1)	(2)	(3)	(4)	(5)	(6)
VARIABLES	Baseline	SVI	Extended	SVI	Total migration	SVI
Log population (o)	1.499***	1.872***	1.220***	1.412***	1.171***	1.501***
	(0.287)	(0.252)	(0.243)	(0.226)	(0.173)	(0.150)
Log income (o)	0.0874	0.0505	0.0413	0.0269	0.0919	0.0638
	(0.0791)	(0.0587)	(0.0783)	(0.0599)	(0.0620)	(0.0453)
House price index			0.000360*	0.000421***		
			(0.000197)	(0.000144)		
Unemployment rate			0.0200***	0.0207***		
			(0.00656)	(0.00590)		
Cost of living index			0.00444***	0.00408***		
			(0.00119)	(0.00118)		
Internet user			-0.170	-0.120		
			(0.164)	(0.157)		
No. days over 90 degree			0.000275	0.000326		
			(0.000340)	(0.000321)		
Political party strength			0.00817	-0.00253		
			(0.0114)	(0.00883)		
SVI (unilateral)		✓		✓		✓
Joint significance SVI		0.000		0.000		0.000
Fixed effects						
Origin	✓	✓	✓	✓	✓	✓
Year	✓	✓	✓	✓	✓	✓
Observations	714	714	714	714	714	714
Observations per predictor	507	11.3	89.3	10.3	507	11.3
R-squared	0.252	0.421	0.342	0.480	0.288	0.443
Within-R^2	0.252	0.421	0.342	0.480	0.333	0.471
Overall-R^2	0.995	0.996	0.996	0.997	0.998	0.999

Notes: Each column displays the results of a separate regression based on equation 1. The dependent variable is the log of the annual migration outflows with a one-year lag from a given origin state to all other states for columns 1-4, while the dependent variable in columns 5 and 6 is the log of the annual migration outflows from a given origin state to the same state and other states.
Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

The estimation of the unilateral migration model with a large vector of the SVI variables raises two issues: multicollinearity and overfitting. First, the SVI variables could be correlated with each other. While the resulting multicollinearity does not exert any detrimental effect on the precision of prediction, it might make it challenging to find a realistic interpretation of the coefficient estimates. Second, the unilateral model with the SVI variables has a small sample size relative to the number of covariates ratio, which may result in overfitting. Harrell (2001) and Babyak (2004) suggest that 10 – 15 observations per regression parameter is the minimum required for a linear regression model to avoid overfitting. Although the unilateral model with the SVI variables has 10-11 observations per predictor, the unilateral model still can create an overfitting problem due to the multicollinearity of the SVI variables.

I address the above problems in two ways (the results are reported in Appendices A through D). First, I conduct a principal component analysis (PCA) to reduce the dimensionality of the data. The PCA consolidates the correlated variables into new covariates that are linear combinations of the original variables. These new predictors, the so-called principal components, are uncorrelated (i.e., orthogonal) to each other. The first six principal components from the SVI variables explain 75% of the variance in the SVI. The within-R^2 in the principal component regression that replaces the SVI variables with six principal components still increases the within-R^2 to 31.4%, compared to 25.2% in column 1. Second, I use the least absolute shrinkage and selection operator (LASSO) to reduce the high variance by selecting the SVI variables. The LASSO method is commonly used for correlated variable selection by using cross-validation to determine both the number of included predictors and the degree of shrinkage to avoid overfitting. The LASSO selects 36 SVI variables (out of 62) with the lowest mean squared prediction error. Lastly, I conduct out-of-sample predictions to test how the PCA and the LASSO

model perform in predicting out-of-sample, using k-fold cross-validation techniques. I find that both the selected SVI variables and principal components mitigating correlation issues also boost the accuracy of the unilateral model from out-of-sample.

4.2 Bilateral model

The results of estimations of various Equation 2 variants are presented in Table 6. I estimate five different baseline fixed-effects models that do not include the bilateral SVI variables but control for a varying set of covariates. To develop the bilateral model specification, I follow the example of Beine et al. (2016) and use the four most common bilateral migration models from the literature (Anderson, 2011; Mayda, 2010; Ortega & Peri, 2013), as well as one recently devised model (Böhme et al., 2020). All these specifications differ with respect to the configurations of fixed effects control variables. Then, I run the same models with the bilateral SVI variables and evaluate the increase in the model fit statistic. I include the home price index, unemployment rate, consumer price index, share of internet users, number of days over 90 degrees, and political party strength for origin and destination state across all models. Similar to the interpretations of unilateral model specification, I focus on the within-R^2 for the bilateral model specifications, which measures how well the explanatory variables account for changes in the bilateral migration flows within the panel unit (origin state) over time.

Columns 1 and 2 report the results from the most basic fixed effects specification of Mayda (2010). This first specification includes the destination, origin, and year fixed effects as well as extended explanatory variables. Almost all variables are statistically significant except for the number of days with temperatures above 90 degrees. The coefficients of the house price index (origin) and the unemployment rate (origin) variables are positive, while the coefficients of

the house price index (destination) and the unemployment rate (destination) variables are negative. These results imply that the increase in house prices in the origin states has a positive effect on outflows from the origin states, while the increase in house prices in the destination states has a negative effect on outflows from the origin states. Unemployment rates have a similar effect to the effect of house prices. The overall-R^2 in column 1 is 71.3% because the fixed effects absorb most of the effects of time-variant and all of the effects of time-invariant factors on migration flows. However, the within-R^2 is only 0.51%. This low within-R^2 suggests that the origin and destination control variables have low predictive power for each side of the migration corridor. However, adding the bilateral SVI variables substantially increases the within-R^2 to 61.0% (column 2), and the overall- R^2 also increases to 88.7%. The bilateral SVI variables are jointly statistically significant, which means that the bilateral SVI variables are associated with bilateral migration flows.

The second fixed effect specification replicates estimations in Ortega and Peri (2013) and includes destination and origin-year fixed effects to capture the effect of the barriers to moving to other destinations (*k*, *l*, etc.) on migration from state *i* to state *j*. All time-invariant destination factors and all time-variant origin factors are absorbed by fixed effects. As a result, all time-variant origin variables cannot be included in the estimations and were removed from the model specifications. The within-R^2 in column 3 is 0.24%, a notable decrease from the results presented in column 1, which is caused by deleting the origin control variables. However, the overall-R^2 increases slightly to 71.4%, compared to 71.3% in column 1. After adding the bilateral SVI variables, the within-R^2 increases to 65.1%, showing a greater improvement relative to the first specification (columns 1 and 2). The bilateral SVI variables are jointly statistically significant.

Table 6: Results of Bilateral Regression

	A		B		C		D		E	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
VARIABLES	Baseline	SVI	Baseline	SVI	Baseline	SVI	Baseline	SVI	Baseline	SVI
Log income (o)	0.0569***	-0.0214								
	(0.0232)	(0.0542)								
Log population (o)	1.2577***	1.0620***								
	(0.0721)	(0.2385)								
Log income (d)	0.1582***	0.0702	0.1581***	0.1161**			0.1582***	0.1180***		
	(0.0260)	(0.0464)	(0.0252)	(0.0421)			(0.0261)	(0.0238)		
Log population (d)	0.9290***	0.9045***	0.9290***	0.7359***			0.9290***	0.8521***		
	(0.0668)	(0.1379)	(0.0614)	(0.1254)			(0.0636)	(0.0634)		
House price index (o)	0.0005***	0.0014***								
	(0.0001)	(0.0002)								
Unemployment rate (o)	0.0220***	0.0307***								
	(0.0017)	(0.0050)								
Cost of living index (o)	0.0044***	0.0072***								
	(0.0005)	(0.011)								
Internet user (o)	0.0279	0.7539***								
	(0.0595)	(0.1522)								
No. days over 90 degree (o)	0.0002***	0.0013***								
	0.0001	(0.0002)								
Political party strength (o)	0.0199***	0.0333***								
	(0.0044)	(0.0101)								
House price index (d)	-0.0005***	-0.0006***	-0.0005***	-0.0006***			-0.0005***	-0.0006***		
	(0.0001)	(0.0001)	(0.0001)	(0.0001)			(0.0001)	(0.0001)		
Unemployment rate (d)	-0.0462***	-0.0320***	-0.0462***	-0.0319**			-0.0462***	-0.0460***		
	(0.0017)	(0.0035)	(0.0015)	(0.0031)			(0.0016)	(0.0016)		
Cost of living index (d)	0.0026***	0.0031***	0.0026***	0.0038***			0.0026***	0.0024***		
	(0.0005)	(0.0009)	(0.0004)	(0.0008)			(0.0004)	(0.0004)		
Internet user (d)	-0.0970*	-0.3381***	-0.0970**	-0.2814***			-0.0970**	-0.0827**		
	(0.0580)	(0.1122)	(0.0532)	(0.1018)			(0.0552)	(0.0549)		
No. days over 90 degree (d)	-0.0001	0.0000	-0.0001	0.0001			-0.0001	-0.0001		
	(0.0001)	(0.0001)	(0.0001)	(0.0001)			(0.0001)	(0.0001)		
Political party strength (d)	0.0435***	0.0400***	0.0435***	0.0387***			0.0435***	0.0404***		

	(0.0051)	(0.0084)	(0.0049)	(0.0077)			(0.0051)	(0.0047)		
SVI (bilateral)		✓		✓		✓		✓		✓
Joint significance SVI		0.0000		0.0000		0.0000		0.0000		0.0000
Fixed Effects										
Destination	✓	✓	✓	✓						
Origin	✓	✓								
Year	✓	✓								
Destination-Year					✓	✓			✓	✓
Origin-Year			✓	✓	✓	✓	✓	✓	✓	✓
Destination-Origin							✓	✓	✓	✓
Observations	35700	35700	35700	35700	35700	35700	35700	35700	35700	35700
Within-R^2	0.0051	0.610	0.0024	0.651	0.000	0.667	0.085	0.104	0.000	0.021
Overall-R^2	0.713	0.887	0.714	0.900	0.716	0.905	0.993	0.993	0.995	0.995

Notes: Each column displays the results of a separate regression based on equation 2. Dependent variable is the log of the annual migration outflows from a given origin state to a specific destination state.
Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1

In the third gravity model with fixed effects from Anderson (2011), origin-year and destination-year fixed effects are added to the model, which captures inward and outward multilateral resistance to migration. All time-variant origin and destination factors are absorbed by fixed effects, and therefore all control variables are removed from the model. As a result, the within-R^2 is zero in column 5, while the overall-R^2 is 71.6%. However, the within-R^2 in the SVI model increases to 66.7%, and the overall-R^2 reaches 90.5% in the full specification. The bilateral SVI variables are jointly statistically significant.

The fourth specification adds the origin-destination and origin-year fixed effects. It is the most demanding specification, as all time-invariant bilateral factors (e.g., state borders) and all time-variant origin factors are absorbed by fixed effects. Similar to the second specification, a set of destination control variables has low predictive power, with the within-R^2 equal to 8.5%. On the other hand, the overall-R^2 increases to 99.3% in column 7 due to the origin-destination pair fixed effects. In column 8, the overall-R^2 remains the same, but the within-R^2 slightly increases to 10.4%. Only destination-year and pair-year variations are left in this model, but the bilateral SVI variables still provide the additional predictive power for bilateral migration flows. The bilateral SVI variables are jointly statistically significant.

In the last specification, fashioned after Böhme et al. (2020), destination-year, origin-year, and destination-origin fixed effects are added to the model. All time-invariant and time-variant bilateral factors are absorbed by fixed effects. The overall-R^2 in columns 9 and 10 is the same, 99.5%. Since almost all between variations are removed, I can observe the SVI variables' sole contribution to the predictive power from the within-R^2 in column 10. The within-R^2 increases from zero to 2.1%. The bilateral SVI variables are jointly statistically significant.

In sum, the results from the five specifications are consistent and confirm that the SVI variables are associated with bilateral migration flows, and provide additional predictive power relative to the mainstream models used in migration literature because all within-R^2 values in the five SVI models are greater than the five baseline models, even though their magnitude decreases in the most demanding specification. Therefore, Hypotheses 1 and 2 are confirmed in the bilateral model specification. In terms of overfitting, bilateral models are less problematic than unilateral models because bilateral models have a large ratio of observations per parameter to prevent overfitting.

4.3 Discussion

In this chapter, two research questions were examined: 1) can search engine data provide information about migration intentions from an origin area? 2) can search engine data improve the predictive power of migration flows forecast models? Two model specifications, unilateral and bilateral, are used to test the determinants of general migration intentions and destination-specific migration intentions. The predictive power of SVI for unilateral and bilateral migration decisions is assessed relative to other determinants.

In the unilateral analysis, the results from three specifications are consistent and point to the joint statistical significance of the unilateral SVI variables and a sizable increase in within-R^2 values. The evidence of associations between unilateral SVI variables and unilateral migration flows provides solid support for Hypothesis 1. The unilateral SVI variables also provide additional predictive power for unilateral migration flows, extending support to Hypothesis 2. The within-R^2 values of SVI models are substantially higher than those of models without the SVI variables.

I found a similar pattern of results in the bilateral analysis. Across the five specifications, the p-values from the joint hypothesis test for the bilateral SVI variables are close to zero (Hypothesis 1) and the within-R^2 values of the SVI models are all higher than the models without the bilateral SVI variables (Hypothesis 2).

It would be informative to investigate the relationship between the SVI variables and directly expressed migration intentions. To do so, I use the Panel Study of Income Dynamics (PSID), a longitudinal nationwide survey of more than 18,000 individuals living in 5,000 families in the United States. The PSID inquiries about moving intentions with the following two questions:

1. *Do you think you might move in the next couple of years?*

   1: Yes, might or maybe
   5: No
   8: DK
   9: NA

2. *Would you say you definitely will move, probably will move, or are you more uncertain?*

   1: Definitely
   2: Probably
   3: More uncertain
   8: DK
   9: NA; refused
   0: Inap: will not move (Q1=5); DK, NA, or RF whether will move (Q1 = 8 or 9)

Only the respondents who answer in the affirmative to the first question are asked the second question. Therefore, I use the first question to identify moving intention and combine the answer with the respondents' current location of residence. I analyze the PSID data for the period that overlaps the analyses of online search volumes presented above. However, since the PSID data in this period are collected biennially, only odd-numbered years corresponding to the

unilateral migration data (2004-2018) can be matched to the previous analyses (e.g., 2005, 2007, 2009, 2011, 2013, 2015, and 2017).

A simple OLS regression is used to test the relationship between the SVI variables and migration intentions. The dependent variable is the log of the number of responses in a state indicating an intention to move, and the independent variables are the set of the SVI variables. Results are reported in Appendix F. The number of observations for this analysis is 357, half of the unilateral model in Table 5, because I have only odd year observations. The coefficients of 18 SVI variables out of 63 have a statistically significant association with migration intention measured in the PSID. Most of these keywords are related to housing and employment (e.g., "buy house," "rent apartment," "job," "work," "salary," etc.). The overall-R^2 is 85.4%, which means that the SVI explains a very high variation of the PSID migration intentions despite the absence of year and origin fixed effects. Overall, this result supports my conclusion that the SVI variables can be used directly to measure migration intentions.

Do policymakers indeed benefit from my methodological approach? I believe the answer is yes. According to the recent study by Gathergood et al. (2021), the COVID-19 pandemic has accelerated the gaps in wealth inequality in the United Kingdom. While the overall consumption in the United Kingdom has declined due to the COVID-19, the wealthier areas such as London's commuter belt have recovered their consumer spending faster than the less-affluent areas. The authors tracked individual spending behavior over time using real-time data on credit card transactions and checking account balances. They found that the variation in these data overlapped with the subsequent variation in official statistics released several months later. Therefore, Gathergood et al. (2021) argued that "the ability to measure regional, economic data in real-time using datasets such as Fable Data offers exciting potential to inform when, where

and how to target regional policy interventions for evidence-based policymaking" (p. 29). That is, the real-time data provide policymakers extra time to adjust current policies. One author (Guttman-Kenney) also mentioned that "with real-time data, we can work out precisely where to target policy and then track its effectiveness to make a decision in weeks whether to modify the measure" (Nelson, 2021, para.9).

The methodological approach used in this dissertation introduces a tool designed to significantly increase the predictive power of models forecasting migration flows without detailed demographic information. Several past studies predicted U.S. interstate migration (e.g., Frees, 1992, 1993; Isserman et al., 1985). These studies predicted U.S. migration rates using the IRS data, and the authors argued that forecasting U.S. migration rates would not be easy because the IRS data provide almost no demographic details. Demographic information has traditionally been considered crucial to the predictions of migration patterns (Isserman et al., 1985). Frees (1993) complained that non-demographic explanatory variables in regressions provide little predictive power for migration rates. Such an observation is consistent with my results in the bilateral migration specifications (Model A). However, this dissertation found that the use of online search data can help achieve high precision in predicting migration flows even without detailed demographic data.

It should be acknowledged, however, that the capability of online search data to improve predictions varies by application. Limnios and You (2021) found that Google Trends data offered only limited improvement in the accuracy of forecast in models of house prices and concluded that online search data did not provide substantial additional predictive power of house prices. The authors used only one search keyword, "[city] Real Estate Agency". Given that house prices are determined by multiple factors, such as property, neighborhood, environmental,

and economic characteristics, the use of a single search keyword might offer an alternative explanation for the low forecasting performance of the Limnios and You's (2021) models.

The purpose of this dissertation is to demonstrate the possible applications of big data obtained from search engines. The prospects for the enrichment of forecast modeling with big data from Google Trends seem attractive for several reasons. First, Google Trends data are freely available. Even though Google imposes a daily limit of downloads for search volume indexes, the limit does not seem too prohibitive for most applications. Moreover, the accessibility of various real-time data will likely increase in the future, presenting new opportunities for researchers to improve the precision of predictions of various outcomes. Second, Google Trends can effectively measure consumers' interests or concerns. In particular, Google Trends data provide valuable insights during crises such as the COVID-19 pandemic. Since it is relatively easy to infer consumer concerns and behavioral intentions from search data, they may provide policymakers with a particularly useful tool in times when the speed of policy decision-making is important. Lastly, as more Americans use search engines, especially Google, the search engine user base will grow more representative of the broader population. According to the Pew Research Center (2012), 91% of online adults (18-65+) use search engines to find information online. Specifically, the 18-20 age group has the highest percentage (96%) of online adults in an age group, followed by 50-64 (92%), 30-49 (91%), and 65+ (80%). Therefore, Google Trends data will become fully representative of the general population.

This dissertation has several limitations. First, I can not claim that my analyses identified causal relationships between the SVI variables and migration flows. Yet, this dissertation provides valuable implications for policymakers to adjust current policies and design new policies related to current events. Second, Google Trends provides normalized data on search

frequencies rather than the measures of absolute numbers of searches. Therefore, it is not possible to compare actual search volumes between search keywords. Third, Google Trends data are random samples of total search term volumes; therefore, Google Trends can provide different search volume indexes even though the same keywords, timelines, and geographic locations are used. In particular, the inconsistent daily Google Trends data can lead to different results. However, as Eichenauer et al. (2021) explained, relying on monthly or annual Google Trends data mitigates this issue.

I have some suggestions for future research to further advance the proposed approach to modeling migration flows as presented in this paper. First, longer time series data can increase the predictive power of the presented models and maximize the accuracy of migration flow estimates. Indeed, larger samples permit a more complex model specification and prevent overfitting. Second, future research can compare my model performance to other models for predicting migration flows to validate the model. Third, since migration patterns differ by age, future research can examine the heterogeneity of forecast separately by age groups. Fourth, an interactive map can show estimates geo-spatially and help policymakers make their decisions on visual information.

CHAPTER 5

CONCLUSION

Some states have recently lost population due to interstate migration (e.g., California and New York), while other states (e.g., Florida and Arizona) have increased in population as they are on the receiving end of migration flows. Population changes have significant effects on state economies, and states need to understand and predict migration trends to attract migrants and retain current residents.

In this dissertation, I demonstrate that online search data can be a useful and effective predictor of interstate migration flows that leads to an improvement of the traditional gravity models of migration. The forecasting results indicate that the changes in search intensity can indeed forecast migration flows and provide additional predictive power over the mainstream models used in migration literature. The SVI variables are jointly significant in all models tested in this dissertation, which means that the SVI variables are associated with migration flows. In terms of the predictive power of the SVI variable, the SVI model in the basic fixed effect specification adds more than 60% variance explained over the baseline model. Even in the most demanding specification, the measure of search volume still improves the accuracy of the model.

This dissertation also found that housing and employment are the main drivers of migration from both unilateral and bilateral analyses, which is consistent with previous studies on internal migration. The results show that origin states with rising house prices experience increasing outflows from the origin states, while the increase in housing prices in destination

states leads to decreasing outflows from the origin states; this is consistent with previous studies on internal migration.

The contributions of this dissertation can be summarized as follows. First, my study develops a tool that can provide a more up-to-date understanding of migration flows in the United States. As discussed earlier, official migration statistics often come with a time lag and therefore could fail to correctly capture the full extent of migration, making informed policy decisions a challenge. My approach offers a novel solution that can mitigate the time lag in official migration data. This tool could, for example, be used for short-term policy prediction exercises as a proactive measure in many other fields. Second, this dissertation investigates the potential use of online search data in the field of U.S. internal migration. The focus of previous research on using big data has been limited to international migration. My balanced panel bilateral migration data with no zero observations can provide unbiased estimates for migration flows, which previous studies could not provide due to limited international migration flow data.

REFERENCES

Agrawal, A., Horton, J., Lacetera, N., & Lyons, E. (2015). *8. Digitization and the Contract Labor Market: A Research Agenda* (pp. 219-256). University of Chicago Press.

Ajzen, I. (1991). The theory of planned behavior. *Organizational behavior and human decision processes, 50*(2), 179-211.

Akerman, A., Gaarder, I., & Mogstad, M. (2015). *The skill complementarity of broadband internet. The Quarterly Journal of Economics, 130*(4), 1781-1824.

Allen, W. B. (1972). An economic derivation of the "Gravity Law" of spatial interaction: a comment on the reply. *Journal of Regional Science, 12*(1), 119-126.

Anderson, J. E. (1979). A theoretical foundation for the gravity equation. *The American Economic Review, 69*(1), 106-116.

Anderson, J. E. (2011). The gravity model. *Annual Review Economics, 3*(1), 133-160.

Andrienko, Y., & Guriev, S. (2004). Determinants of interregional mobility in Russia. *Economics of Transition, 12*(1), 1-27.

Askitas, N., & Zimmermann, K. F. (2009). Google econometrics and unemployment forecasting. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1480251

Asur, S., & Huberman, B. A. (2010, August). Predicting the future with social media. In *2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology (Vol. 1, pp. 492-499)*. IEEE.

Babyak, M. A. (2004). What you see may not be what you get: a brief, nontechnical introduction to overfitting in regression-type models. *Psychosomatic medicine, 66*(3), 411-421.

Banbura, M., Giannone, D., & Reichlin, L. (2010). Nowcasting. Retrieved from: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1717887

Banerjee, B. (1984). Information flow, expectations and job search: Rural-to-urban migration process in India. *Journal of Development Economics*, *15*(1-3), 239-257.

Banerjee, B. (1991). The determinants of migrating with a pre-arranged job and of the initial duration of urban unemployment: An analysis based on Indian data on rural-to-urban migrants. *Journal of Development Economics, 36*(2), 337-351.

Bawakid, A., & Oussalah, M. (2010, September). Using features extracted from Wikipedia for the task of word sense disambiguation. *2010 IEEE 9th International Conference on Cyberntic Intelligent Systems* (pp. 1-6). IEEE.

Beine, M., Bertoli, S., & Fernández-Huertas Moraga, J. (2016). A practitioners' guide to gravity models of international migration. *The World Economy, 39*(4), 496-512.

Beracha, E., & Wintoki, M. B. (2013). Forecasting residential real estate price changes from online search activity. *Journal of Real Estate Research, 35*(3), 283-312.

Beringer, G., Jabtloński, M., Januszewski, P., Sobecki, A., & Szymański, J. (2019, September). Towards semantic-rich word embeddings. *2019 Federated Conference on Computer Science and Information Systems (FedCSIS)* (pp. 273-276). IEEE.

Bethlehem, J., & Biffignandi, S. (2012). Handbook of web surveys. Hoboken. NJ: Wiley & Sons. Couper, M.(1997). Survey introductions and data quality. *Public Opinion Quarterly, 61*(2), 317-338.

Böheim, R., & Taylor, M. P. (2002). Tied down or room to move? Investigating the relationships between housing tenure, employment status and residential mobility in Britain. *Scottish Journal of Political Economy, 49*(4), 369-392.

Böhme, M. H., Gröger, A., & Stöhr, T. (2020). Searching for a better life: Predicting international migration with online search keywords. *Journal of Development Economics, 142*, 102347.

Boustan, L. P., Fishback, P. V., & Kantor, S. (2010). The effect of internal migration on local labor markets: American cities during the Great Depression. *Journal of Labor Economics, 28*(4), 719-746.

Bradlow, E. T., & Schmittlein, D. C. (2000). The little engines that could: Modeling the performance of World Wide Web search engines. *Marketing Science, 19*(1), 43-62.

Burger, M., Van Oort, F., & Linders, G. J. (2009). On the specification of the gravity model of trade: zeros, excess zeros and zero-inflated estimation. *Spatial Economic Analysis, 4*(2), 167-190.

Carneiro, H. A., & Mylonakis, E. (2009). Google trends: a web-based tool for real-time surveillance of disease outbreaks. *Clinical Infectious Diseases, 49*(10), 1557-1564.

Carr, C. T., & Hayes, R. A. (2015). Social media: Defining, developing, and divining. *Atlantic Journal of Communication, 23*(1), 46-65.

Carrière-Swallow, Y., & Labbé, F. (2013). Nowcasting with Google Trends in an emerging market. *Journal of Forecasting, 32*(4), 289-298.

Champion, A. G., Fotheringham, S., Rees, P., Boyle, P., & Stillwell, J. (1998). *The determinants of migration flows in England: A review of existing data and evidence: A report prepared*

for the Department of the Environment, Transport and the Regions. Newcastle upon Tyne: Department of Geography, University of Newcastle upon Tyne.

Chen, J., Chen, H., Wu, Z., Hu, D., & Pan, J. Z. (2017). Forecasting smog-related health hazard based on social media and physical sensor. *Information Systems, 64,* 281-291.

Chen, Y., Jeon, G. Y., & Kim, Y. M. (2014). A day without a search engine: an experimental study of online and offline searches. *Experimental Economics, 17*(4), 512-536.

Choe, J. I. (2003). Do foreign direct investment and gross domestic investment promote economic growth?. *Review of Development Economics*, *7*(1), 44-57.

Choi, C. (2003). Does the Internet stimulate inward foreign direct investment?. *Journal of Policy Modeling*, *25*(4), 319-326.

Choi, H. and Varian, H. (2009a). Predicting the present with Google Trends. Google Inc. Retrieved from: http://google.com/googleblogs/pdfs/google_predicting_the_present.pdf

Choi, H., & Varian, H. (2009b). Predicting initial claims for unemployment benefits. Google Inc. Retrieved from: http://research.google.com/archive/papers/initialclaimsUS.pdf

Choi, H., & Varian, H. (2012). Predicting the present with Google Trends. *Economic Record*, *88*, 2-9.

Choo, S., & Mokhtarian, P. L. (2007). Telecommunications and travel demand and supply: Aggregate structural equation models for the US. *Transportation Research Part A: Policy and Practice*, *41*(1), 4-18.

Chort, I. (2012). New insights into the selection process of Mexican migrants. What can we learn from discrepancies between intentions to migrate and actual moves to the US?. Retrieved from https://halshs.archives-ouvertes.fr/halshs-00689467/document

Cole, J. I., Suman, M., Schramm, P., Lunn, R., & Aquino, J. S. (2003). The UCLA Internet report surveying the digital future: Year three. Retrieved from http://www.digitalcenter.org/wp-content/uploads/2013/02/2003_digital_future_report-year3.pdf

Constantiou, I. D., & Kallinikos, J. (2015). New games, new rules: big data and the changing context of strategy. *Journal of Information Technology, 30*(1), 44-57.

Conway, D. (1980). Step-wise migration: Toward a clarification of the mechanism. *International Migration Review, 14*(1), 3-14.

Cooke, T. J. (2013). Internal migration in decline. *The Professional Geographer, 65*(4), 664-675.

Czernich, N., Falck, O., Kretschmer, T., & Woessmann, L. (2011). Broadband infrastructure and economic growth. *The Economic Journal, 121*(552), 505-532.

D'Amuri, F., & Marcucci, J. (2017). The predictive power of Google searches in forecasting US unemployment. *International Journal of Forecasting, 33*(4), 801-816.

Da, Z., Engelberg, J., & Gao, P. (2015). The sum of all FEARS investor sentiment and asset prices. *The Review of Financial Studies, 28*(1), 1-32.

DaVanzo, J. (1978). Does unemployment affect migration? Evidence from micro data. *The Review of Economics and Statistics*, 504-514.

DaVanzo, J. (1983). Repeat migration in the United States: who moves back and who moves on?. *The Review of Economics and Statistics*, 552-559.

Daveri, F., & Faini, R. (1999). Where do migrants go?. *Oxford Economic Papers, 51*(4), 595-622.

de Haas, H. (2011). The Determinants of International Migration: Conceptualizing Policy, Origin and Destination Effects. Oxford: International Migration Institute. IMI Working Paper No. 32.

de Haas, H. (2021). A theory of migration: the aspirations-capabilities framework. *Comparative Migration Studies, 9*(1), 1-35.

De Jong, G. F. (2000). Expectations, gender, and norms in migration decision-making. *Population Studies, 54*(3), 307-319.

De Jong, G. F., Root, B. D., Gardner, R. W., Fawcett, J. T., & Abad, R. G. (1985). Migration intentions and behavior: Decision making in a rural Philippine province. *Population and Environment, 8*(1-2), 41-62.

Ding, W., & Marchionini, G. (1996). A comparative study of web search service performance. Proceedings of the ASIST Annual Meeting, 33, 136. Retrieved from https://www.learntechlib.org/p/83946/

Dong, X., & Su, L. T. (1997). Search engines on the World Wide Web and information retrieval from the Internet: A review and evaluation. *Online and CD-ROM Review, 21*(2), 67-82.

Dorigo, G., & Tobler, W. (1983). Push-pull migration laws. *Annals of the Association of American Geographers, 73*(1), 1-17.

Dowling, G. R., & Staelin, R. (1994). A model of perceived risk and intended risk-handling activity. *Journal of Consumer Research, 21*(1), 119–134. doi: 10.1086/209386.

Eichenauer, V. Z., Indergand, R., Martínez, I. Z., & Sax, C. (2022). Obtaining consistent time series from Google Trends. *Economic Inquiry, 60*(2), 694-705.

Elder, G. H., Johnson, M. K., & Crosnoe, R. (2003). *The emergence and development of life course theory*. In Handbook of the life course (pp. 3-19). Springer, Boston, MA.

Elgendy, N., & Elragal, A. (2016). Big data analytics in support of the decision making process. *Procedia Computer Science*, 100, 1071-1084.

Ettredge, M., Gerdes, J., & Karuga, G. (2005). Using web-based search data to predict macroeconomic statistics. *Communications of the ACM, 48*(11), 87-92.

Farhoodi, M., Mahmoudi, M., Bidoki, A. Z., Yari, A., & Azadnia, M. (2009). Query expansion using persian ontology derived from Wikipedia. *World Applied Sciences Journal, 7*(4), 410-417.

Ferreira, F., Gyourko, J., & Tracy, J. (2010). Housing busts and household mobility. *Journal of Urban Economics, 68*(1), 34-45.

Fiorio, L., Abel, G., Cai, J., Zagheni, E., Weber, I., & Vinué, G. (2017, June). Using Twitter data to estimate the relationship between short-term mobility and long-term migration. Proceedings of the 2017 ACM on web science conference (pp. 103-110).

Ford, G. T., Smith, D. B., & Swasy, J. L. (1988). An empirical test of the search, experience and credence attributes framework. ACR North American Advances (pp.239-244).

Fouberg, E. H., Murphy, A. B., & De Blij, H. J. (2015). *Human geography: people, place, and culture*. John Wiley & Sons.

Frees, E. W. (1992). Forecasting state-to-state migration rates. *Journal of Business & Economic Statistics, 10*(2), 153-167.

Frees, E. W. (1993). *Short-term forecasting of internal migration. Environment and Planning A, 25*(11), 1593-1606.

Frey, W. (2009). *The great American migration slowdown*. Brookings Institution, Washington, DC.

Frey, W. H. (2020, December 15). Just before covid-19, American migration hit a 73-year low. Brookings. Retrieved April 18, 2022, from https://www.brookings.edu/blog/the-avenue/2020/12/15/just-before-covid-19-american-migration-hit-a-73-year-low/

Frost, R. (2020). Are Americans Stuck in Place? Declining Residential Mobility in the US. Joint Center for Housing Studies of Harvard University. Retrieved from https://www.jchs.harvard.edu/sites/default/files/harvard_jchs_are_americans_stuck_in_place_frost_2020.pdf

Furceri, D. (2006). Does labour respond to cyclical fluctuations? The case of Italy. *Applied Economics Letters*, *13*(3), 135-139.

Gathergood, J., Gunzinger, F., Guttman-Kenney, B., Quispe-Torreblanca, E., & Stewart, N. (2021). Levelling down and the covid-19 lockdowns: Uneven regional recovery in uk consumer spending. Covid Economics, 67, 24-52.

Gayo-Avello, D. (2012). No, you cannot predict elections with Twitter. *IEEE Internet Computing*, *16*(6), 91-94.

Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., & Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, *457*(7232), 1012-1014.

Gödri, I., & Feleky, G. A. (2017). Selection of migrants and realization of migration intentions-lessons from a panel study (No. 29). Working papers on population, family and welfare. Retrieved from https://www.econstor.eu/handle/10419/226472

Graves, P. E. (1983). Migration with a composite amenity: the role of rents. *Journal of Regional Science*, *23*(4), 541-546.

Graves, P. E., & Linneman, P. D. (1979). Household migration: Theoretical and empirical results. *Journal of Urban Economics, 6*(3), 383-404.

Greenwood, M. J. (1997). *Internal migration in developed countries*. Handbook of population and family economics, 1, 647-720.

Grigg, D. B. (1977). EG Ravenstein and the "laws of migration". *Journal of Historical Geography*, *3*(1), 41-54.

Gross, E. (2005). Internal revenue service area-to-area migration data: strengths, limitations, and current uses. *Statistics of Income. SOI Bulletin, 25*(3), 159-160.

Guo, C. (2001), "A review on consumer external search: amount and determinants", *Journal of Business and Psychology, 15*, 505-19.

Gupta, A., & Das, P. (2022). Asymmetric political attention across foreign and domestic private equity real estate investors. *Journal of Property Research, 39*(1), 1-29.

Harrell, F. E. (2001). *Regression modeling strategies: with applications to linear models, logistic regression and survival analysis*. New York: Springer; 2001. 4. McCullagh P,

Harris, J. R., & Todaro, M. P. (1970). Migration, unemployment and development: a two-sector analysis. *The American Economic review*, 126-142.

Hauser, J. R., Urban, G. L., & Weinberg, B. D. (1993). How consumers allocate their time when searching for information. *Journal of Marketing Research, 30*(4), 452-466.

Hecht, B., & Stephens, M. (2014, May). A tale of cities: Urban biases in volunteered geographic information. Proceedings of the International AAAI Conference on Web and Social Media (Vol. 8, No. 1).

Heckman, J. J. (1979). Sample selection bias as a specification error. *Econometrica: Journal of the Econometric Society*, 153-161.

Helpman, E., Melitz, M., & Rubinstein, Y. (2008). Estimating trade flows: Trading partners and trading volumes. *The Quarterly Journal of Economics, 123*(2), 441-487.

Henley, A. (1998). Residential mobility, housing equity and the labour market. *The economic journal, 108*(447), 414-427.

Hiller, N., & Lerbs, O. W. (2016). Aging and urban house prices. *Regional Science and Urban Economics, 60*, 276-291.

Hudson, S., Huang, L., Roth, M. S., & Madden, T. J. (2016). The influence of social media interactions on consumer–brand relationships: A three-country study of brand perceptions and marketing behaviors. *International Journal of Research in Marketing, 33*(1), 27-41.

Iftikhar, R., & Khan, M. S. (2020). Social media big data analytics for demand forecasting: development and case implementation of an innovative framework. *Journal of Global Information Management (JGIM), 28*(1), 103-120.

Isard, W. (1960). *Methods of regional science*. Cambridge, Mass.

Isotalo, V., Saari, P., Paasivaara, M., Steineker, A., & Gloor, P. A. (2016). Predicting 2016 US presidential election polls with online and media variables. In *Designing Networks for Innovation and Improvisation* (pp. 45-53). Springer, Cham.

Isserman, A. M., Plane, D. A., Rogerson, P. A., & Beaumont, P. M. (1985). Forecasting interstate migration with limited data: A demographic-economic approach. *Journal of the American Statistical Association, 80*(390), 277-285.

Jackman, R., & Savouri, S. (1992). Regional migration versus regional commuting: the identification of housing and employment flows. *Scottish Journal of Political Economy, 39*(3), 272-287.

Jansen, B. J., & Molina, P. R. (2006). The effectiveness of Web search engines for retrieving relevant ecommerce links. *Information Processing & Management, 42*(4), 1075-1098.

Jansen, B. J., Booth, D. L., & Spink, A. (2008). Determining the informational, navigational, and transactional intent of Web queries. *Information Processing & Management, 44*(3), 1251-1266.

Jansen, B. J., Spink, A., & Pedersen, J. (2005). A temporal comparison of AltaVista Web searching. *Journal of the American Society for Information Science and Technology, 56*(6), 559-570.

Johnes, G., & Hyclak, T. (1994). House prices, migration, and regional labor markets. *Journal of Housing Economics, 3*(4), 312-329.

Joseph, K., Wintoki, M. B., & Zhang, Z. (2011). Forecasting abnormal stock returns and trading volume using investor sentiment: Evidence from online search. *International Journal of Forecasting, 27*(4), 1116-1127.

Kaplan, G., & Schulhofer-Wohl, S. (2017). Understanding the long-run decline in interstate migration. *International Economic Review, 58*(1), 57-94.

Kau, J. B., & Sirmans, C. F. (1979). The functional form of the gravity model. *International Regional Science Review, 4*(2), 127-136.

Keikha, A., Ensan, F., & Bagheri, E. (2018). Query expansion using pseudo relevance feedback on wikipedia. *Journal of Intelligent Information Systems, 50*(3), 455-478.

Kley, S. (2011). Explaining the stages of migration within a life-course framework. *European Sociological Review, 27*(4), 469-486.

Kley, S. A., & Mulder, C. H. (2010). Considering, planning, and realizing migration in early adulthood. The influence of life-course events and perceived opportunities on leaving the city in Germany. *Journal of Housing and the Built Environment, 25*(1), 73-94.

Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2009). Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery, 18*(1), 140-181.

Kotorri, M., Krasniqi, B. A., & Dabic, M. (2020). Migration, remittances, and entrepreneurship: a seemingly unrelated bivariate probit approach. *Remittances review, 5*(1), 15-36.

Krugman, P. (1980). Scale economies, product differentiation, and the pattern of trade. *The American Economic Review, 70*(5), 950-959.

Kumar, N., & Lang, K. R. (2007). Do search terms matter for online consumers? The interplay between search engine query specification and topical organization. *Decision Support Systems, 44*(1), 159-174.

Kumar, N., Lang, K. R., & Peng, Q. (2005, January). Consumer search behavior in online shopping environments. *Proceedings of the 38th Annual Hawaii International Conference on System Sciences (pp. 175b-175b).* IEEE.

Kuruzovich, J., Viswanathan, S., Agarwal, R., Gosain, S., & Weitzman, S. (2008). Marketspace or marketplace? Online information search and channel outcomes in auto retailing. *Information Systems Research, 19*(2), 182-201.

Lampos, V., & Cristianini, N. (2012). Nowcasting events from the social web with statistical learning. *ACM Transactions on Intelligent Systems and Technology (TIST), 3*(4), 1-22.

Larkin, M. P., Askarov, Z., Doucouliagos, H., Dubelaar, C., Klona, M., Newton, J., Stanley, T.D., & Vocino, A. (2019). Do house prices ride the wave of immigration?. *Journal of Housing Economics, 46*, 101630.

Lee, E. S. (1966). A theory of migration. *Demography, 3*(1), 47-57.

Lee, B. K., & Lee, W. N. (2004). The effect of information overload on consumer choice quality in an on-line environment. *Psychology & Marketing, 21*(3), 159-183.

Limnios, A. C., & You, H. (2021). Can Google Trends Improve Housing Market Forecasts?. *Curiosity: Interdisciplinary Journal of Research and Innovation, 1*(2), 21987.

Lin, A. Y., Cranshaw, J., & Counts, S. (2019, May). Forecasting us domestic migration using internet search queries. The World Wide Web Conference (pp. 1061-1072).

Lin, Y., Ma, Z., Zhao, K., Hu, W., & Wei, J. (2018). The impact of population migration on urban housing prices: Evidence from China's major cities. *Sustainability, 10*(9), 3169.

Lu, M. (1998). Analyzing migration decisionmaking: Relationships between residential satisfaction, mobility intentions, and moving behavior. *Environment and Planning A, 30*(8), 1473-1495.

Malik, M. M., Lamba, H., Nakos, C., & Pfeffer, J. (2015, April). Population bias in geotagged tweets. Ninth international AAAI conference on web and social media.

Mavragani, A., Ochoa, G., & Tsagarakis, K. P. (2018). Assessing the methods, tools, and statistical approaches in Google Trends research: systematic review. *Journal of Medical Internet Research*, *20*(11), e270.

Mayda, A. M. (2010). International migration: A panel data analysis of the determinants of bilateral flows. *Journal of Population Economics, 23*(4), 1249-1274.

Mayda, A. M., Peri, G., & Steingress, W. (2018). The political impact of immigration: Evidence from the United States (No. w24510). National Bureau of Economic Research.

Meshcheryakov, A. (2018). Using online search queries in real estate research with an empirical example of arson forecast. *Journal of Real Estate Literature, 26*(2), 331-361.

Mincer, J. (1978). Family migration decisions. *Journal of Political Economy, 86*(5), 749-773.

Molloy, R., Smith, C. L., & Wozniak, A. (2011). Internal migration in the United States. *Journal of Economic Perspectives, 25*(3), 173-96.

Molloy, R., Smith, C. L., & Wozniak, A. (2017). Job changing and the decline in long-distance migration in the United States. *Demography, 54*(2), 631-653.

Nelson, B. (2021, February 19). How real-time data can help target policy in a pandemic. The University of Chicago Booth School of Business. Retrieved April 18, 2022, from https://www.chicagobooth.edu/review/how-real-time-data-can-help-target-policy-pandemic

Nelson, P. (1970). Information and consumer behavior. *Journal of Political Economy, 78*(2), 311-329.

Niedomysl, T. (2011). How migration motives change over migration distance: Evidence on variation across socio-economic and demographic groups. *Regional Studies, 45*(6), 843-855.

Nilsson, K. (2003). Moving into the city and moving out again: Swedish evidence from the cohort born in 1968. *Urban Studies, 40*(7), 1243-1258.

Oktay, H., Taylor, B. J., & Jensen, D. D. (2010, July). Causal discovery in social media using quasi-experimental designs. In *Proceedings of the first workshop on social media analytics* (pp. 1-9).

Orcle. What is Big Data?. Retrieved April 18, 2022, from https://www.oracle.com/big-data/what-is-big-data/

Ormerod, P., Nyman, R., & Bentley, R. A. (2014). Nowcasting economic and social data: when and why search engine data fails, an illustration using Google Flu Trends. arXiv preprint arXiv:1408.0699.

Ortega, F., & Peri, G. (2013). The effect of income and immigration policies on international migration. *Migration Studies, 1*(1), 47-74.

Pagolu, V. S., Reddy, K. N., Panda, G., & Majhi, B. (2016, October). Sentiment analysis of Twitter data for predicting stock market movements. *In 2016 international conference on signal processing, communication, power and embedded system (SCOPES) (pp. 1345-1350).* IEEE.

Parikh, A., & Van Leuvensteijn, M. (2003). Internal migration in regions of Germany: A panel data analysis. *Applied Economics Quarterly*, *49*(2), 173-192.

Peterson, R. A., & Merino, M. C. (2003). Consumer information search behavior and the Internet. *Psychology & Marketing*, *20*(2), 99-121.

Pew Research Center. (2021, November 23). *Demographics of internet and home broadband usage in the United States*. Pew Research Center: Internet, Science & Tech. Retrieved March 31, 2022, from https://www.pewresearch.org/internet/fact-sheet/internet-broadband/?menuItem=d5edf003-5858-4269-89c5-f2889ecf7951

Piras, R. (2021). Migration flows by educational attainment: Disentangling the heterogeneous role of push and pull factors. *Journal of Regional Science, 61*(3), 515-542.

Preis, T., Moat, H. S., & Stanley, H. E. (2013). Quantifying trading behavior in financial markets using Google Trends. *Scientific Reports*, *3*(1), 1-6.

Purcell, K., Brenner, J., & Rainie, L. (2012). *Search engine use 2012*. Pew research center's internet & American life project.

Ratchford, B. T. (1982). Cost-benefit models for explaining consumer choice and information seeking behavior. *Management Science, 28*(2), 197-212.

Ratchford, B. T., Lee, M. S., & Talukdar, D. (2003). The impact of the Internet on information search for automobiles. *Journal of Marketing research, 40*(2), 193-209.

Ravenstein, E. G. (1885). The laws of migration. *Journal of the Statistical Society of London, 48*(2), 167-235.

Ravenstein, E. G. (1889). The laws of migration. *Journal of the Royal Statistical Society, 52*(2), 241-305.

Renas, S. M., & Kumar, R. (1978). The cost of living, labor market opportunities, and the migration decision: A case of misspecification?. *The Annals of Regional Science, 12*(2), 95-104.

Riddell, J. B., & Harvey, M. E. (1972). The urban system in the migration process: an evaluation of step-wise migration in Sierra Leone. *Economic Geography, 48*(3), 270-283.

Roback, J. (1988). Wages, rents, and amenities: differences among workers and regions. Economic Inquiry, 26(1), 23-41.

Rossi, C., Acerbo, F. S., Ylinen, K., Juga, I., Nurmi, P., Bosca, A., Tarasconi, F., Cristoforetti, M., & Alikadic, A. (2018). Early detection and information extraction for weather-induced floods using social media streams. *International journal of disaster risk reduction, 30*, 145-157.

Rousidis, D., Koukaras, P., & Tjortjis, C. (2020). Social media prediction: a literature review. *Multimedia Tools and Applications, 79*(9), 6279-6311.

Sadilek, A., Kautz, H., DiPrete, L., Labus, B., Portman, E., Teitel, J., & Silenzio, V. (2016, March). Deploying nEmesis: Preventing foodborne illness by data mining social media. In *Twenty-Eighth IAAI Conference.*

Saks, R. E., & Wozniak, A. (2011). Labor reallocation over the business cycle: New evidence from internal migration. *Journal of Labor Economics, 29*(4), 697-739.

Schachter, J. (2001). Why people move: Exploring the March 2000 current population survey (Vol. 204). US Department of Commerce, Economics and Statistics Administration, Bureau of the Census.

Schachter, J. (2004). *Geographical Mobility, 2002-2003*. Washington, DC: US Department of Commerce, Economics and Statistics Administration, Bureau of the Census.

Scharkow, M., & Vogelgesang, J. (2011). Measuring the public agenda using search engine queries. *International Journal of Public Opinion Research, 23*(1), 104-113.

Schoen, H., Gayo-Avello, D., Metaxas, P. T., Mustafaraj, E., Strohmaier, M., & Gloor, P. (2013). The power of prediction with social media. *Internet Research*.

Schonlau, M., Van Soest, A., Kapteyn, A., & Couper, M. (2009). Selection bias in web surveys and the use of propensity scores. *Sociological Methods & Research, 37*(3), 291-318.

Schwarzer, M., Schubotz, M., Meuschke, N., Breitinger, C., Markl, V., & Gipp, B. (2016, June). Evaluating link-based recommendations for Wikipedia. *2016 IEEE/ACM Joint Conference on Digital Libraries (JCDL)* (pp. 191-200). IEEE.

Silva, J. S., & Tenreyro, S. (2006). The log of gravity. *The Review of Economics and statistics, 88*(4), 641-658.

Skeldon, R. (1990). *Population mobility in developing countries*. Belhaven Press.

Speare, A. (1974). Residential satisfaction as an intervening variable in residential mobility. *Demography, 11*(2), 173-188.

State, B., Rodriguez, M., Helbing, D., & Zagheni, E. (2014). Migration of Professionals to the US. Social Informatics: 6th International Conference, SocInfo 2014, Barcelona, Spain, November 11–13, 2014. Proceedings, Lecture Notes in Computer Science (pp. 531-43). Cham, Switz: Springer.

State, B., Weber, I., & Zagheni, E. (2013, February). Studying inter-national mobility through ip geolocation. Proceedings of the sixth ACM international conference on Web search and data mining (pp. 265-274).

Stawarz, N., Sander, N., & Sulak, H. (2021). Internal migration and housing costs—A panel analysis for Germany. *Population, Space and Place*, *27*(4), e2412.

Stigler, G. J. (1961). The economics of information. *Journal of Political Economy*, *69*(3), 213-225.

Su, L. T. (1998). Value of search results as a whole as the best single measure of information retrieval performance. *Information Processing & Management, 34*(5), 557-579.

Subramani, S., Michalska, S., Wang, H., Whittaker, F., & Heyward, B. (2018, October). Text mining and real-time analytics of twitter data: A case study of australian hay fever prediction. In *International Conference on Health Information Science* (pp. 134-145). Springer, Cham.

Tinbergen, J. (1962). Shaping the world economy. Twentieth Century Fund, New York, NY.; suggestions for an international economic policy.

U.S. Census Bureau. (2021, April). Computer and Internet Use in the United States: 2018. Census.gov. Retrieved April 18, 2022, from https://www.census.gov/content/dam/Census/library/publications/2021/acs/acs-49.pdf

U.S. Census Bureau. (2021, December 3). About migration and place of birth. Census.gov. Retrieved April 18, 2022, from https://www.census.gov/topics/population/migration/about.html

Van Dalen, H. P., & Henkens, K. (2013). Explaining emigration intentions and behaviour in the Netherlands, 2005–10. *Population Studies, 67*(2), 225-241.

Van Hear, N., Bakewell, O., & Long, K. (2018). Push-pull plus: reconsidering the drivers of migration. *Journal of ethnic and migration studies, 44*(6), 927-944.

Vanderkamp, J. (1971). Migration flows, their determinants and the effects of return migration. *Journal of Political Economy, 79*(5), 1012-1031.

Varian, H. R. (2014). Big data: New tricks for econometrics. *Journal of Economic Perspectives, 28*(2), 3-28.

Vidal, M., Menezes, G. V., Berlt, K., de Moura, E. S., Okada, K., Ziviani, N., Fernandes, D., & Cristo, M. (2012, October). Selecting keywords to represent web pages using wikipedia information. *Proceedings of the 18th Brazilian Symposium on Multimedia and the Web* (pp. 375-382).

Vilhelmson, B., & Thulin, E. (2013). Does the Internet encourage people to move? Investigating Swedish young adults' internal migration experiences and plans. *Geoforum, 47*, 209-216.

Wanner, P. (2021). How well can we estimate immigration trends using Google data?. *Quality & Quantity, 55*(4), 1181-1202.

Winkler, H. (2017). How does the internet affect migration decisions?. *Applied Economics Letters, 24*(16), 1194-1198.

Winkler, R. L., & Rouleau, M. D. (2021). Amenities or disamenities? Estimating the impacts of extreme heat and wildfire on domestic US migration. *Population and Environment, 42*(4), 622-648.

Winship, C., & Mare, R. D. (1992). Models for sample selection bias. *Annual Review of Sociology, 18*(1), 327-350.

Wu, L., & Brynjolfsson, E. (2015). *3. The Future of Prediction: How Google Searches Foreshadow Housing Prices and Sales*. Economic analysis of the digital economy (pp. 89-118). University of Chicago Press.

Wu, Z., Zhu, H., Li, G., Cui, Z., Huang, H., Li, J., Chen, E., & Xu, G. (2017). An efficient Wikipedia semantic matching approach to text document classification. *Information Sciences, 393*, 15-28.

Yin, Z. H., & Choi, C. H. (2021). Has the Internet increased FDI, economic growth, and trade? Evidence from Asian economies. *Information Development*.

Zabel, J. E. (2012). Migration, housing market, and labor market responses to employment shocks. *Journal of Urban Economics, 72*(2-3), 267-284.

Zagheni, E., & Weber, I. (2012, June). You are where you e-mail: using e-mail data to estimate international migration rates. Proceedings of the 4th annual ACM web science conference (pp. 348-351).

Zagheni, E., & Weber, I. (2015). Demographic research with non-representative internet data. *International Journal of Manpower*.

Zagheni, E., Weber, I., & Gummadi, K. (2017). Leveraging Facebook's advertising platform to monitor stocks of migrants. *Population and Development Review*, 721-734.

Zagheni, E., Garimella, V. R. K., Weber, I., & State, B. (2014, April). Inferring international and internal migration patterns from twitter data. Proceedings of the 23rd International Conference on World Wide Web (pp. 439-444).

Zamani, M., & Schwartz, H. A. (2017, April). Using twitter language to predict the real estate market. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*: Volume 2, Short Papers (pp. 28-33).

Zhang, Y. (2006, August). Wiki means more: hyperreading in Wikipedia. *Proceedings of the seventeenth conference on Hypertext and hypermedia* (pp. 23-26).

Zhao, F., Fang, F., Yan, F., Jin, H., & Zhang, Q. (2012). Expanding approach to information retrieval using semantic similarity analysis based on WordNet and Wikipedia. *International Journal of Software Engineering and Knowledge Engineering*, *22*(02), 305-322.

Appendix A: Principal Component Analysis of Google Trends Index variables

Component	Eigenvalue	Difference	Proportion	Cumulative
Comp1	21.9562	7.43873	0.3599	0.3599
Comp2	14.5175	9.48648	0.238	0.5979
Comp3	5.03097	3.34899	0.0825	0.6804
Comp4	1.68198	0.185834	0.0276	0.708
Comp5	1.49615	0.421815	0.0245	0.7325
Comp6	1.07433	0.110422	0.0176	0.7501
Comp7	0.963908	0.0461135	0.0158	0.7659
Comp8	0.917795	0.0820309	0.015	0.781
Comp9	0.835764	0.0478224	0.0137	0.7947
Comp10	0.787942	0.091104	0.0129	0.8076

Notes: Appendix A shows the eigenvalues for each component, the difference in eigenvalue size between the components, the proportion of variation explained by each component, and the cumulative proportion explained. I choose the first six components because their eigenvalues are greater than one. The first six components explain 75% of the variation in the SVI variables.

Appendix B: Out-Of-Sample Exercise with Principal Component Analysis

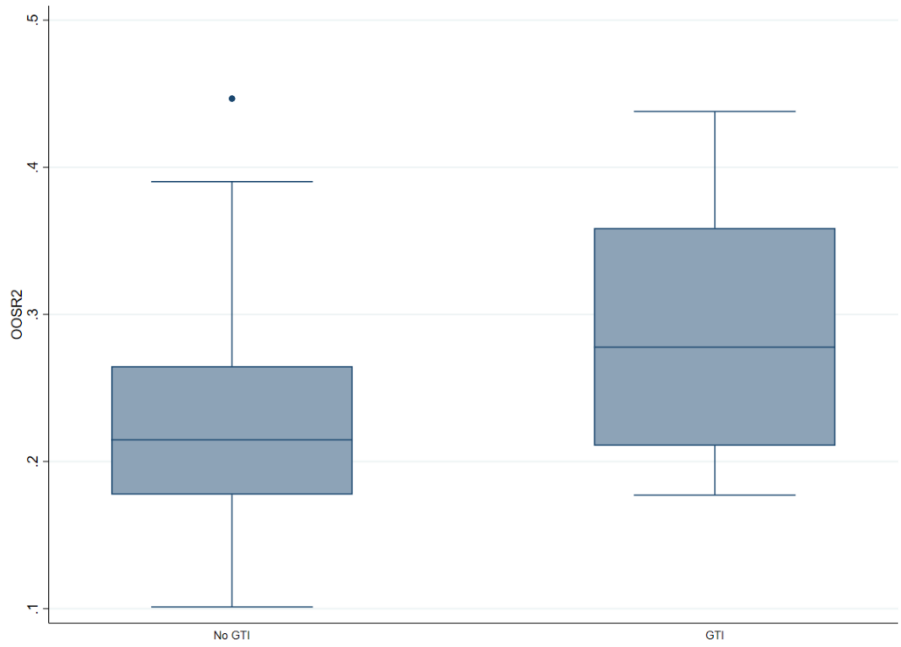
	RMSE	Std. Err.	[95% Conf.	Interval]
Model (pc1)	0.216	0.004	0.206	0.226
Model (pc1-pc2)	0.206	0.006	0.193	0.220
Model (pc1-pc3)	0.201	0.007	0.186	0.217
Model (pc1-pc4)	0.197	0.006	0.183	0.210
Model (pc1-pc5)	0.194	0.006	0.180	0.209
Model (pc1-pc6)	0.189	0.007	0.173	0.205
Model (pc1-pc7)	0.190	0.006	0.176	0.205

Notes: This table shows the RMSE for each regression. To find the best predictors, I calculated the estimated RMSE for a model with the first principal component from the SVI variables as a predictor with k=10. Then, I repeat the process with the first and second principal components as predictors and continue adding principal components to the model until RMSE does not decrease significantly. I use the function seven times, once for each model. It confirms that the RMSE was the lowest when six principal components are used as predictors in the out-of-sample observations.

Appendix C: Principal Component Regression

	A		B	
	(1)	(2)	(3)	(4)
VARIABLES	Baseline	SVI	Extended	SVI
Log population (origin)	1.499***	1.542***	1.220***	1.252***
	(0.287)	(0.290)	(0.243)	(0.254)
Log income (origin)	0.0874	0.0936	0.0413	0.0666
	(0.0791)	(0.0740)	(0.0783)	(0.0745)
House price index			0.000360*	0.000366**
			(0.000197)	(0.000174)
Unemployment rate			0.0200***	0.0239***
			(0.00656)	(0.00687)
Cost of living index			0.00444***	0.00431***
			(0.00119)	(0.00112)
Internet user			-0.170	-0.190
			(0.164)	(0.171)
No. days over 90 degree			0.000275	0.000292
			(0.000340)	(0.000349)
Political party strength			0.00817	0.00496
			(0.0114)	(0.0108)
pc1		0.00913*		0.00405
		(0.00533)		(0.00481)
pc2		0.00136		0.00874
		(0.00750)		(0.00604)
pc3		-0.00338		0.00516
		(0.00463)		(0.00353)
pc4		0.00129		-0.00691
		(0.00597)		(0.00675)
pc5		-0.00229		0.000565
		(0.00893)		(0.00767)
pc6		0.00102		0.00324
		(0.00666)		(0.00602)
Joint significance PC		0.241		0.139
Fixed effects				
Origin	✓	✓	✓	✓
Year	✓	✓	✓	✓
Observations	714	714	714	714
Within-R^2	0.252	0.270	0.342	0.358
Overall-R^2	0.995	0.995	0.996	0.996

Notes: Appendix C shows the results of PCA regression with six principal components in in-sample observation. Even though the within-R^2 in the SVI model (column 2) is higher than the baseline model, the joint significance of six principal components is 0.241, and is not statistically significant. Therefore, PCA is insufficient to handle multicollinearity. I observe a similar effect in the extended model (columns 3 and 4), with the increased within-R^2 and insignificance of principal components.
Robust standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1



Notes: The same baseline model in Table 5 is used with the selected SVI variables. LASSO selects 36 out of 60 SVI variables, and the p-value of the test of joint significance for 36 SVI variables is less than 0.001. Then, I compare the out-of-sample within-R^2. Appendix D shows that the SVI model has s higher within-R^2 than the baseline model, which explains higher variance than the baseline model. Overall, the SVI model performs well in the out-of-sample observations and is not caused by overfitting.

Appendix D: Out-Of-Sample within-R^2 of LASSO Regression

Appendix E: Unilateral regression with the PSID variable

VARIABLES	(1) PSID
Category: real estate	-0.00650
	(0.00924)
Category: real estate agencies	0.00200
	(0.00549)
Category: real estate listings	0.0106
	(0.00697)
Category: apartments & residential rentals	-0.00205
	(0.00519)
Category: jobs	-0.0139
	(0.00856)
Category: developer jobs	0.0254***
	(0.00332)
Category: jobs & education	0.00614
	(0.00969)
Category: job listings	0.0102**
	(0.00469)
Keyword: housing	0.0156***
	(0.00541)
Keyword: house	0.000220
	(0.00929)
Keyword: apartment	0.00794
	(0.00698)
Keyword: real estate	0.00814
	(0.00703)
Keyword: property	-0.00747
	(0.00505)
Keyword: relocation	0.0110*
	(0.00631)
Keyword: moving	0.00569
	(0.00568)
Keyword: buy house	0.0144***
	(0.00500)
Keyword: renthouse	0.00649
	(0.00469)
Keyword: rent apartment	-0.0115**
	(0.00555)
Keyword: mortgage	0.00597
	(0.00719)
Keyword: Zillow	-0.00632
	(0.00522)
Keyword: redfin	-0.00109
	(0.00327)
Keyword: Trulia	-0.00191
	(0.00474)
Keyword: realtor	0.00570
	(0.00448)

Keyword: Remax	0.00236
	(0.00539)
Keyword: cost of living	0.0178**
	(0.00711)
Keyword: home	-0.0173***
	(0.00651)
Keyword: neighborhood	0.0179***
	(0.00446)
Keyword: metropolitan	-0.0137**
	(0.00565)
Keyword: city	0.00555
	(0.00471)
Keyword: employment	-0.000701
	(0.00669)
Keyword: job	0.0243***
	(0.00579)
Keyword: work	0.0177**
	(0.00815)
Keyword: occupation	-0.00203
	(0.00623)
Keyword: payroll	-0.00190
	(0.00538)
Keyword: minimum wage	0.0110***
	(0.00417)
Keyword: unemployment	0.00126
	(0.00406)
Keyword: internship	0.00339
	(0.00480)
Keyword: career	-0.00371
	(0.00538)
Keyword: labor	0.00630
	(0.00496)
Keyword: layoff	0.00855
	(0.00574)
Keyword: employer	-0.00301
	(0.00483)
Keyword: hiring	0.00519
	(0.00499)
Keyword: income	0.00750
	(0.00666)
Keyword: salary	0.0192***
	(0.00653)
Keyword: recruitment	-0.0114**
	(0.00495)
Keyword: welfare	-0.00932*
	(0.00488)
Keyword: rush hour	0.00443
	(0.00629)
Keyword: traffic	-0.00278
	(0.00405)

Keyword: climate	-0.00766
	(0.00504)
Keyword: tax	-0.0371***
	(0.00845)
Keyword: state income tax	-0.00407
	(0.00857)
Keyword: weather	0.00747
	(0.00589)
Keyword: temperature	-0.0153**
	(0.00718)
Keyword: crime	0.00153
	(0.00533)
Keyword: school	-0.00342
	(0.00757)
Keyword: health	-0.00491
	(0.00680)
Keyword: foreclosure	0.00481
	(0.00507)
Keyword: commute	7.07e-05
	(0.00533)
Keyword: retired	0.00138
	(0.00496)
Keyword: college	0.00651
	(0.00801)
Keyword: university	-0.00416
	(0.00587)
Constant	-0.513
	(0.858)
Observations	357
R-squared	0.854

Notes: The dependent variable is the log of migration intentions from the PSID, and the independent variables are the set of the SVI variables.
Standard errors in parentheses
*** p<0.01, ** p<0.05, * p<0.1