Dimensions of Interstate and Intrastate Household Migration, 1990-2015

Michael Andrew Babb

A dissertation
submitted in partial fulfillment of the
requirements for the degree of

Doctor of Philosophy

University of Washington

2021

Reading Committee:

John Mark Ellis, Chair

Suzanne D Withers

Michael P. Brown

Program Authorized to Offer Degree:
Geography

University of Washington

Abstract

Dimensions of Interstate and Intrastate Household Migration, 1990-2015

Michael Andrew Babb

Chair of the Supervisory Committee:
Professor John Mark Ellis
Department of Geography

This dissertation enhances and makes use of county-to-county migration data to investigate several dimensions of interstate and intrastate migration during the 1990 through 2015 period. The first empirical chapter describes an algorithm that combines spatial interaction modelling and linear optimization programming techniques to enhance publicly available county-to-county migration data. The algorithm produces estimates of flows between county-to-county pairs not previously reported due to censoring. The implementation of this algorithm generates an additional approximately 500K records per year: tripling the within-state migration complete case count and increasing the interstate migration complete case count by 13-fold. The increased records mean that the migration of an additional approximately 1.6 million households per year can be incorporated into subsequent analyses. The inclusion of additional records serves to decrease potential biases originating from the omission of data. The second empirical chapter uses the enhanced county-to-county migration data to fit production constrained, origin specific, spatial interaction models illustrating the spatiality of households' destination preferences and trends over time. Households in the Great Plains and the South selected destinations with relatively larger populations while households in western states opted for destinations with relatively smaller populations. In general, households in eastern states moved shorter distances while households in western states moved longer distances. The preference for in-state destinations varies by state: Cal-

ifornia, Florida and portions of Washington, Texas, and Ohio showed a nearly consistent preference for out-of-state destinations while large swaths of states in the Great Plains and Rocky Mountain regions exhibited preferences for in-state destinations. The third and final empirical chapter examines the determinants of household movement using a combination of consistent county boundaries, enhanced county-to-county household migration flows, and place-based characteristics such as the age composition of the population, economic, and amenity indicators. County-to-county movement is classified into one of four categories based on the metropolitan statuses of each origin and destination county-to-county pair. The results show how origin and destination place-based characteristics promote outgoing household flows, attract incoming household flows, how these change over time and according to the metropolitan statuses of the origin county and the destination county. Young adults ages 20 through 29 demonstrate the most consistent push and pull factors while other age groups exhibit a variety of patterns over time. Adults ages 60-69 attract households, though this varies by the metropolitan status of the origin and destination counties. Frequently, but not always, an age groups' ability to promote the outgoing migration of households or attract incoming households reached its maximum or minimum around 2000 with a smaller inflection during the late 2000s. Households are generally attracted to counties with higher annual average pay, more affordable non-metropolitan counties, and non-metropolitan counties with high amenities.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACKNOWLEDGMENTS

Without a doubt, this dissertation has been quite a journey. Ten plus years, some blood, some sweat, and certainly some tears. It wasn't always easy, it wasn't always fun, but this section, however, has been the most fulfilling portion to write. My favorite part about graduate school is the people I have met and gotten to know along the way. I've traveled the US and the world because of these relationships and my life is infinitely more rich for this experience and the people I have met during this time.

This dissertation began in the Spring of 2013 in Professor Ellis' iteration of Geography 542: Research Seminar in Social and Population Geography. Professor Ellis' vision for the class was to fit spatial interaction models on a data set that required minimal cleaning. We turned to the IRS County to County migration data. In my mind that class was meant to be a fun project exploring several classic techniques in population geography and migration modelling. That seminar facilitated my growth as a researcher, a scholar, and most importantly, a computer programmer.

I wish to express sincere appreciation to Professor Ellis for the introduction to this topic and his extreme patience and encouragement in seeing this dissertation through. (And a good portion of the other dissertation!)

I thank Professor Suzanne Withers for her ultra-keen methodological and statistical insights. My conversations with Professor Withers always left me in better spirits and helped me be a more thoughtful resesarcher.

I wish to thank Professor Michael Brown for introducing me to geography and his encouragement to see the linkages between seemingly disparate fields of thought. Wednesday, October 23, 2003, around 10:00 in the morning in Professor Brown's Geography 100 class is

when I knew what I was going to do with my life: be a geographer. It was Guggenheim Hall. I was wearing my off-white jeans, a blue polo shirt, and a courderoy jacket. I left that class with the crystal clear realization that I knew what I was going to be doing with the rest of my life. Thank you

To Dr. Jerry Herting: thank you for hiring me at CSSCR at a much needed time and for putting me in touch with who would ultimately become Maximum Analytics' first client.

To Professor Bill Beyers, thank you for believing in me and your encouragement and ideas.

I thank CSDE. This dissertation would not exist without CSDE. Several individuals in particular. Thank you to Dr. Matt Dunbar for being so supportive. He is among the rarefied few yet many for believing in me. To Matt Weatherford and Alan Li: you guys are computing wizards. Thank you for the flawless infrastructure. Phil Hurvitz! You helped me keep my cool throughout this. I owe you some oysters!

To Dr. Tim Thomas: thank you for believing in me and keeping me focused and excited about this project. I can only begin to offer my gratitude for helping me complete this. The next glass of rye will always be on me.

To Dr. Kelly Walsh: thank you for the encouragement, a lifetime of adventures and memories, and being the incredible inspiration you are.

To Chelsey MacNeill: thank you for being such an incredible badass. Thank you for being you.

To Dr. Michalis Avraam: thanks buddy! You really know your stuff. Looking forward to the next adventure. As always.

To Dr. Todd Faubion: thank you for the years of friendship and leading by example.

To Dr. Laine Rutledge: you are fantastic. Rarely does one meet someone with such an effortless and vast intellect. Thank you for the delightful conversations.

To Dr. Austin Gross: thank you for being the most energetic and cynical person I have

ever met. You are, as ever, the most perfect economist. It's been a delight.

To Dr. Becky Burnett: thank you for so many fun nights and wonderful conversations.

To Dr. Tricia Ruiz: thank you for the support and encouragement. You helped me a lot over the years. Thank you.

To Dr. Brandon Derman: thank you! I have always enjoyed talking with you. You are quite the inspiration.

To Dr. Chris Lizotte: thank you for the multiple rounds of encouragement. You helped me keep going.

To Dr. Autumn Knowlton: thank you for the years of friendship! And the support. I am grateful for our friendship.

To Dr. Mónica Farías: thank you for the support and your perseverance. It is an inspiration.

To Dr. Tiffany Grobelski: thank you for the support.

To Dr. Will Buckingham: for showing me it can be done and the years of friendship. And the delightful conversations over text during what would otherwise have been a lonely time.

Thank you David S. Moore for helping me get this started.

Thank you Des Ropel for helping me keep this going.

I wish to thank Matt Townley for helping me see myself and life in general a little more clearly. We've walked a strange and unique path: here's to the next one.

I am grateful to Dr. Agnieszka Leszcynski for the time we shared together.

I thank Dr. Skye Naslund for the happy memories, support, and thrilling travels.

To David Goggins for helping me realize that I am a baddass motherfucker. Stay hard.

To Tina Tian, thank you for always having my back.

To Dr. Cori Mar: thank you for the strong words of encouragement and sage advice.

To Scott Sipes: you made a lot of things look easy. Thanks!

To Dr. Cynthia Brewer: although we've never met, thank you for the colors. Data look a lot better because of you.

To Marshall Agnew: I loved talking about technology with you. Thank you for encouraging me to keep at it.

To my Federal Way boys: Scott, Rollie, Sam, Oliver, and Jeff. Thank you for being a part of this.

To Dr. Uncle Tony: Thank you for the encouragement!

To Sandy and Gordon: Thank you for the support! And asking when I would be done. That was fantastic.

To my parents, Walt and Margie: Thank you for the support and encouragement over many years of study.

To my brother Steve: thank you for always listening and the years of encouragement. I finally did it!

And to Jennifer Vogel. Thank you for your optimistic encouragement, unwavering support, fantastic smile, and love. It's been an absolute blast!

As this is a dissertation in geography I would be remiss to not mention the spaces and places that helped facilitate the writing of this dissertation. Whether the space was a refuge from writing or conducive to it, or sometimes a little of both, it is a part of this story.

To start, first and foremost there is my desk in Smith 430. I occupied that desk for 14 years. Thank you and goodbye.

My desk in Savery 116. So comforting!

My home in Wallingford, Seattle. Thank you being the refuge.

Cafe Umbria in Ballard and in Pioneer Square. Thank you for the coffee and the water and the hours and hours I was able to spend writing.

Cafe Zig Zag. Your most excellent cocktails helped brighten spirits and lift my mood and make the future seem within reach.

Pho Thy Thy. Thank you for the many bowls of noodles. Always so nourishing.

Thanks to Cafe Miir for being a nice place to step into when stepping out of my own place.

The Sloop and your Big Beers. Thank you so much!

The George and Dragon, Murphy's and Shawn O'Donnell's: Thank you for the Guinness and the Jameson.

The Starbucks on the 40th of the Columbia Tower. Thanks for the view!

El Diablo for being the right space at the write time.

El Camion: life's always better after one of your for tacos.

The Lava Lounge: nothing quite like a cold Mai Tai after a dizzying writing spell.

36 Stone: Lots of late nights hunched over my laptop. Thanks!

And the Tin Hat. Thank you. You have been invaluable. I look forward to drinking a glass of rye with my dear friend Dr. Tim Thomas without whom this project would not be possible.

To the Census Bureau: thank you for the clearance, a peak into the inner workings of statistical enumeration, and the opportunity to set up a federal statistical research data center. That was pretty cool.

This dissertation featured a great deal of technical work. And I am very happy about that! But a craftsman is only as good as his tools. I have been fortunate in that those tools have opened doors and shown me new worlds. I'm grateful for laptop3 and studio3 for the years of stellar service, my first iPad and GoodReader for expediting knowledge acquisition, Microsoft Access for teaching me SQL, SQL Server for showing me what it can do, and Postgres for showing me what it can be, csde-ts3 for always being there, SAS for being such a pain in the ass that I quit it, R for being just hard enough that I wanted to learn and love it more, python for being so easy and slick that it made the simple instantaneous and the difficult trivial, and VBA for being my first programming language.

To ArcGIS and ESRI, thank you for showing me so much and helping me show others vast realms of geographic information. Thank you for being so difficult, so damn unreliable, so late to the game, and so nakedly profit driven that I vowed to do everything with open source. Any update on transparent legend graphics?

To Patrick Moore, thank you for your incredible generosity. I would not be be here without you.

To John Kulju, thank you for the internship and the start of this journey.

To Chris Mefford for introducing me to databases, demonstrating that more work is better than less work, deadlines and deliverables are good, and how important it is to keep cool and get it all done.

To Microsoft: thanks for a very interesting job at the right time in my life.

Finally, I wish to express my gratitude to the people of the United States of America for moving about, paying their taxes, and filling out official government forms and the Internal Revenue Service for collecting those taxes and making these data available. Thanks!

DEDICATION

To me for having the stubborn bravery and foolish tenacity to see this through.

Chapter 1

INTERNAL US MIGRATION IN A TIME OF DECLINING MIGRATION RATES

This dissertation is about extending the usability of publicly available US county-to-county household migration data from the Internal Revenue Service (IRS) and the spatiality of internal US migration as measured from the IRS data during the 1990 through 2015 period. A period generally considered to feature declining migration rates. As will be shown in subsequent chapters however, declining migration rates are not the sole facet of internal US migration. Through curation, enhancement, and use of the IRS's county-to-county household migration data, I describe the destination preferences of US households and the demographic and socioeconomic factors influencing moves between, to, and from metropolitan and non-metropolitan counties.

1.1 Overview

The three empirical chapters in this dissertation examine three different aspects of household migration using IRS county-to-county household migration data. In chapter two, I estimate the censored origins and destinations of the publicly available county-to-county migration data. In any given year in the 1990 through 2015 period, approximately 20-to-25-percent of county-to-county household migration flows are censored to protect confidentially and privacy. Flows are either origin censored (a specific destination is known but a specific origin is not) or destination censored (a specific origin is known, but a specific destination is not). Through a combination of spatial interaction modelling and linear optimization programming techniques, I estimate the censored origins and destinations in the county-to-county household migration data. Through this estimation process I triple the number

of completed within-state county-to-county household migration records and I increase the completed interstate county-to-county household migration records by 13-fold. Combining the reported and estimated IRS county-to-county household migration data generates the enhanced county-to-county household migration dataset.

This enhanced county-to-county household migration data is used in chapters three and four. In chapter three, I examine the destination preferences of households during the 1990 through 2015 period using origin-specific, production-constrained spatial interaction models. In practice this means that a regression model is fit for each county for each year in the study period generating coefficients specific to each county. The results from this chapter illustrate the geographic distribution of household preferences for populated destinations, accessible destinations, destination distance, and within-state movement. In chapter four, the final empirical chapter, I model the determinants of four different types of county-to-county household movement using a variety of origin and destination characteristics such as population age structure, affordability, and several natural amenities. This modelling approach emphasizes the differences between movement to, from, and between metropolitan and non-metropolitan counties. I conclude this dissertation in chapter five with recommendations for future study and observations on internal migration in a time of national quarantine brought about by the SARS-CoV-2 coronavirus pandemic beginning in 2020.

1.2 Internal migration in decline

The internal migration analyses carried out in this dissertation use data pertaining to an approximately 30-year period - a period in which scholars generally agree features declining migration rates (Cooke 2013; Molloy et al. 2011). The IRS county-to-county household migration data feature migration rates that sometimes agree with other data sources and at times disagree with other data sources. Because of the rich spatial and temporal detail in the IRS county-to-county household migration data, I can examine the spatialities and temporal trends of internal migration in the US and therefore contribute to larger discussions on the data used to study internal US migration and interstate and intrastate household

movement. Describing the internal migration regime in the United States during the past 30 years underscores that what is known about internal migration is a function of how internal migration is measured - survey versus in-direct measurement - and the scale at which internal migration is reported: within-state versus interstate. Accordingly, getting the data "correct" is a necessary first step in telling the more-than-declining-rates-of-internal-migration story. However, describing the trends in internal migration, as seen from other data sources, highlights the differences in each data source and situates the internal migration narrative and the reason for using the IRS county-to-county household migration data.

Using a variety of data sources, but mostly the Current Population Survey (CPS), Molloy et al. (2011) document the decline in both shorter distance migration rates and longer distance migration rates since the 1980s. Although the authors examine the decline in the migration rate along multiple dimensions, no specific cause is singled out. Migration rates peaked after World War II and have generally been declining. Figure 1.1, page 5, features the counts and rates of migrating individuals in the US from 1968 through 2018 as seen in data from the Current Population Survey (Flood et al. 2018). Figure 1.1 features the number of migrants, graphic A, the total migration rate, graphic B, and individual migration rates, graphic C, from within the same county of origin, moves from within the same state of origin, moves originating in a different state, and moves from foreign locations during the 1968 through 2018 period. Graphic A in Figure 1.1 features the absolute count of incoming migrants by origin status, graphic B expresses those numbers as a combined rate, and graphic C features the migration rate by movement type. Graphic B in Figure 1.1 shows that beginning in 1968, the migration rate was upwards of 20-percent and gradually decreased to approximately 10-percent in 2018. Graphic C shows the general decline in all migration rates. The largest share of migration in any given year is within-county followed by within-state migration. The greatest percentage of people moving reached a peak in 1968 with approximately 35 million people changing homes, about 18-percent of the population. A similar peak was again reached in 1987 with about 42 million people changing homes accounting for 18-percent of the population. Since 1987, both the absolute number of people changing homes and the

relative percent of the population changing homes has decreased; 2017 was the last year in which more than 10-percent of households changed addresses. The within-county migration rate, as depicted in graphic C, shows the greatest decline. The within-state and interstate migration rates show more stability and even feature moderate increases. The CPS data show the within-state migration rate to increase from approximately 3.0-percent in 1968 to almost 4.0-percent in 1987 and then gradually decrease to approximately 2.0-percent in 2018. Aside for a few years in the late 1960s and the early 1970s, the interstate migration rate is always less than the within-state migration rate. The interstate migration rate decreased from a high of 3.5-percent in 1968 to less than 2.0-percent in 2018. The CPS data show that the foreign migration rate is always less than 1.0-percent during the 50-year period.

1.3 Differences in data, differences in migration rates

This dissertation uses county-to-county household migration data from the Internal Revenue Service (IRS) during the 1990 through the 2015 period (Gross 2009, 2005; Internal Revenue Service 2018; Pierce 2015). The IRS data are administrative data originating from the tax-paying households in the US and form a near-population sized dataset covering 95-percent to 98-percent of the individual filing population (Gross 2005). The IRS data report tax returns, a proxy for households, and exemptions, a proxy for individuals. The county-to-county household migration rate is actually the county-to-county tax return migration rate and the county-to-county individual migration rate is actually the county-to-county tax exemption rate. In general, the number of exemptions in any given year is approximately 80-percent of the US population (based on my calculations - see Figure 3.3 on page 90 for an examination of tax returns and occupied households). This approximately 80-percent value is due to not all households filling a tax return: the elderly and those not meeting certain income thresholds). The CPS uses a probabilistic sampling scheme to sample over 65K households in the US (Flood et al. 2018). The differences in data provenance and collection between the CPS and IRS account for the differences in the scale at which publicly available migration data are available in addition to the differences in migration rates.

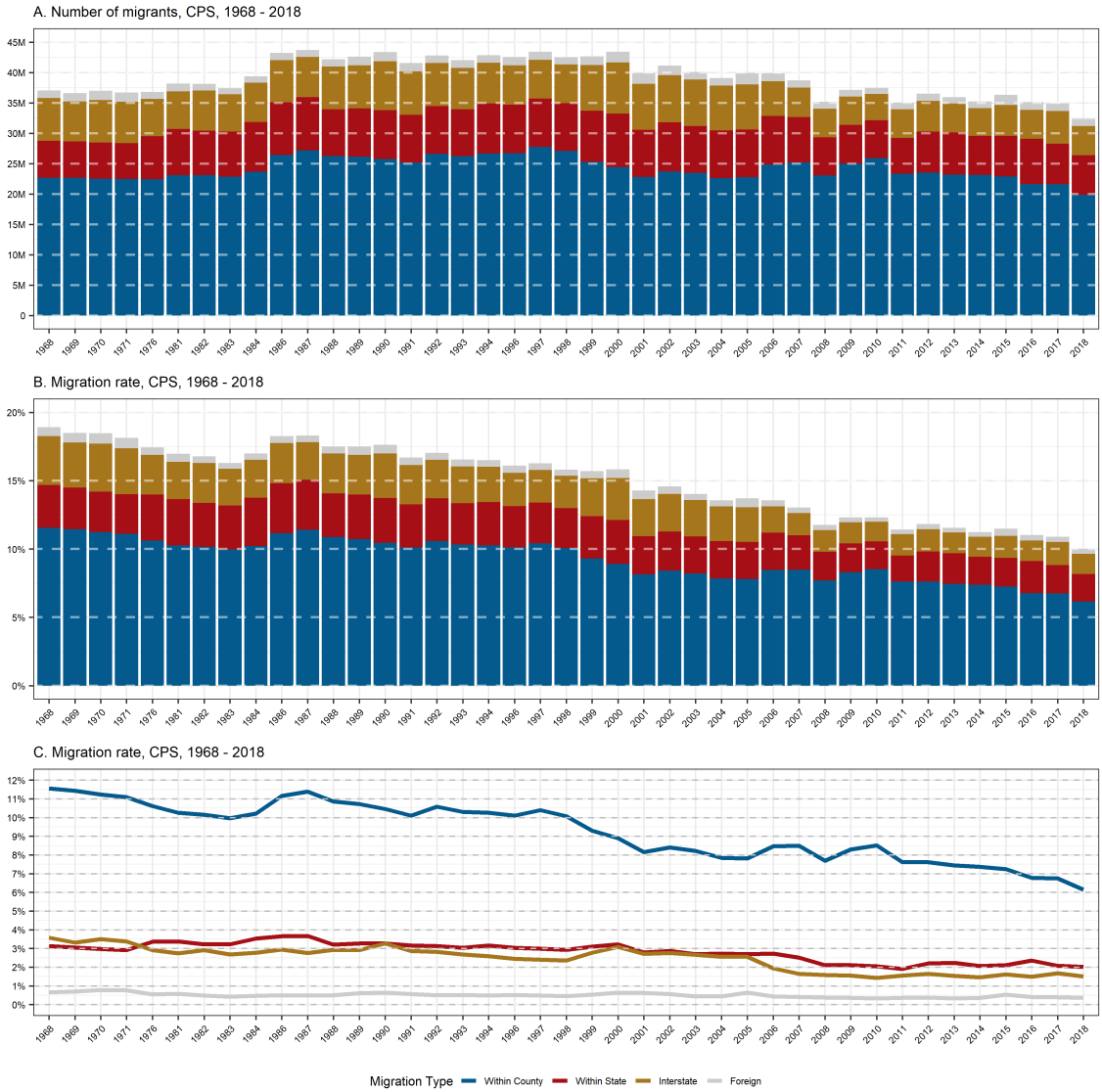


Figure 1.1: Within-county, within-state, interstate, and foreign migration counts and rates, CPS, 1968-2018

The data from the IRS do not feature within-county movement. Comparing the CPS migration data with the IRS migration data requires removing the within-county moves and limiting the data to the 1990 through 2015 period. The three graphics in Figure 1.2 on page 9 feature the counts and rates of within-state, interstate, and foreign migration as depicted in the CPS and IRS data. Note that the data for 1995 has been removed as the CPS data for 1995 is not considered trustworthy (Flood et al. 2018). Graphic A in Figure 1.2 shows the counts of migrating individuals as depicted in the CPS data and graphic B features the counts of migrating individuals as depicted in the IRS data. As shown in graphic A, from 1990 through 2006, the data from the CPS featured a greater number of migrating individuals with almost 19M people migrating in 2000 alone. A noticeable and continued decline is first seen in 2006 and continues through 2010. By 2011, more individuals began migrating. Beginning in 2007 and through 2015, except for 2014, the data from the IRS show greater numbers of migrants. The IRS data as shown in graphic B illustrate increasing numbers of migrants from 1990 through 2005. From 2006 through 2009, the yearly numbers of migrants decreased, on account of the great recession, and by 2010 those numbers were increasing again. In 2013 and 2014, the numbers of migrants decreased and in 2015 the number of migrants was increasing. The dip in 2014 is because of changes the IRS made to data processing (to combat identity theft) and not an actual drop in migration counts (Internal Revenue Service 2018; Pierce 2020). Chapter two discusses the impact of this data processing change and other data processing changes in greater detail. In 2015, the number of migrants increased.

Graphic C in Figure 1.2 features the within-state, interstate, and foreign migration rates as seen in both datasets. In both the CPS data and the IRS data, the within-state migration rate is always greater than the interstate migration rate. The foreign migration rate as shown in the CPS data is always greater than the foreign migration rate as shown in the IRS data. In 1990, the CPS interstate migration rate, the CPS within-state migration rate, and the IRS within-state migration are all about 3.25-percent while the IRS interstate migration rate is about 2.75-percent. After 1990, the migration rates diverge. The IRS migration rates show

less of a decline when compared to the CPS migration rates with the CPS data showing increases in the late 1990s. By 2000, as depicted in the CPS data, both the within-state and interstate migration rates were decreasing and the interstate migration rate, after 2005, began decreasing more quickly than the within-state migration rate. Kaplan and Schulhofer-Wohl (2012) find that the decrease is due to a statistical artifact created by then-current imputation procedures and not because of an actual decrease in the migration rate. When correcting for the change in imputation procedures, the decrease in the migration rate in the CPS data is less sharp. The corrected data, however, are not available from the CPS. Throughout the 1990 through 2015 period, the CPS interstate migration rate is sometimes greater than the IRS interstate migration rate and sometimes less than the CPS interstate migration rate. In 2015, the CPS data feature within-state and interstate migrations rates of approximately 2.1-percent and 1.6-percent, respectively. The CPS intestate migration rate in 2015 was half of what it was in 1990 and the CPS within-state migration rate in 2015 is approximately two-thirds of its value in 1990. Comparing 1990 and 2015 migration rates in the IRS data, the within-state and interstate migration rates feature proportions of a percentage-point of change. In general, the IRS data show less of a decline in the interstate and within-state migration rates when compared to the CPS data.

The differences in migration rates between the CPS and the IRS data are a result of different data collection practices at work. The CPS is a survey and the IRS data are the byproduct of the US's yearly tax collection effort. Not every household in the US pays taxes - some fail to do so and some are exempt - and not every household pays its taxes on time. The data from the IRS are disseminated at a single point in time and not updated when new information is made available, even though the data cover approximately 98-percent of the tax-filing population, about 80-percent of the total population. The migration rate featured in the CPS data is the migration rate of all individuals and the migration rate featured in the IRS data is calculated as the migration rate of tax-filing households and therefore the rates are going to be greater and more consistent. In addition, Molloy et al. (2011, fig. 2) compare migration data from the IRS and the CPS from 1990 to 2010 and migration data

from the American Community Survey from 2005 to 2010 at four different scales: inter-region, interstate, inter-state, and inter-county. Data from the CPS and the ACS show a general decrease over the 30-year period and the IRS data often depict greater rates than those found in the other two data sources.

Focusing on declining internal migration rates is important. However, it obfuscates the fact that people are still moving. As seen in graphic B in figure 1.2, page 9, the data from the IRS show that in 1990, approximately 12.1 million people, 6.1-percent, moved to and around the US. These numbers grew steadily over the 26-year period to reach a value of 14.5 million people in 2015, 5.5-percent. Certainly, millions of people have been and are on the move in the United States and this dissertation investigates the destination preferences of those on the move and why households move. Of course, examining the preferences of household movement is predicated on selecting an appropriate dataset and understanding the intricacies of the data in use.

1.4 Enhancing IRS migration data

What is known about internal migration is often a function of the data sources in use and the search for suitable data on migration is not new and has been an ongoing project over the past forty years. Bilsborrow and Akin (1982) and Isserman et al. (1982) describe the data needs for studying internal migration and the availability of federal survey data for satisfying those data needs. Most research on internal migration was and still is conducted using federal survey data while only recently other data sources have begun to supplement federal survey data. For example, twitter data have been used to study international and internal migration (Zagheni et al. 2014). DeWaard et al. (2018) assess the usefulness of the Consumer Credit Panel, researchers at Microsoft use internet search queries to forecast migration (Lin et al. 2019), and moving companies offer reports on internal migration (Atlas Van Lines 2020; North American Van Lines 2020). In addition to publicly available federal data sources, there are restricted access federal data sources. Restricted access federal data come from the same surveys that generate the publicly available versions. The difference

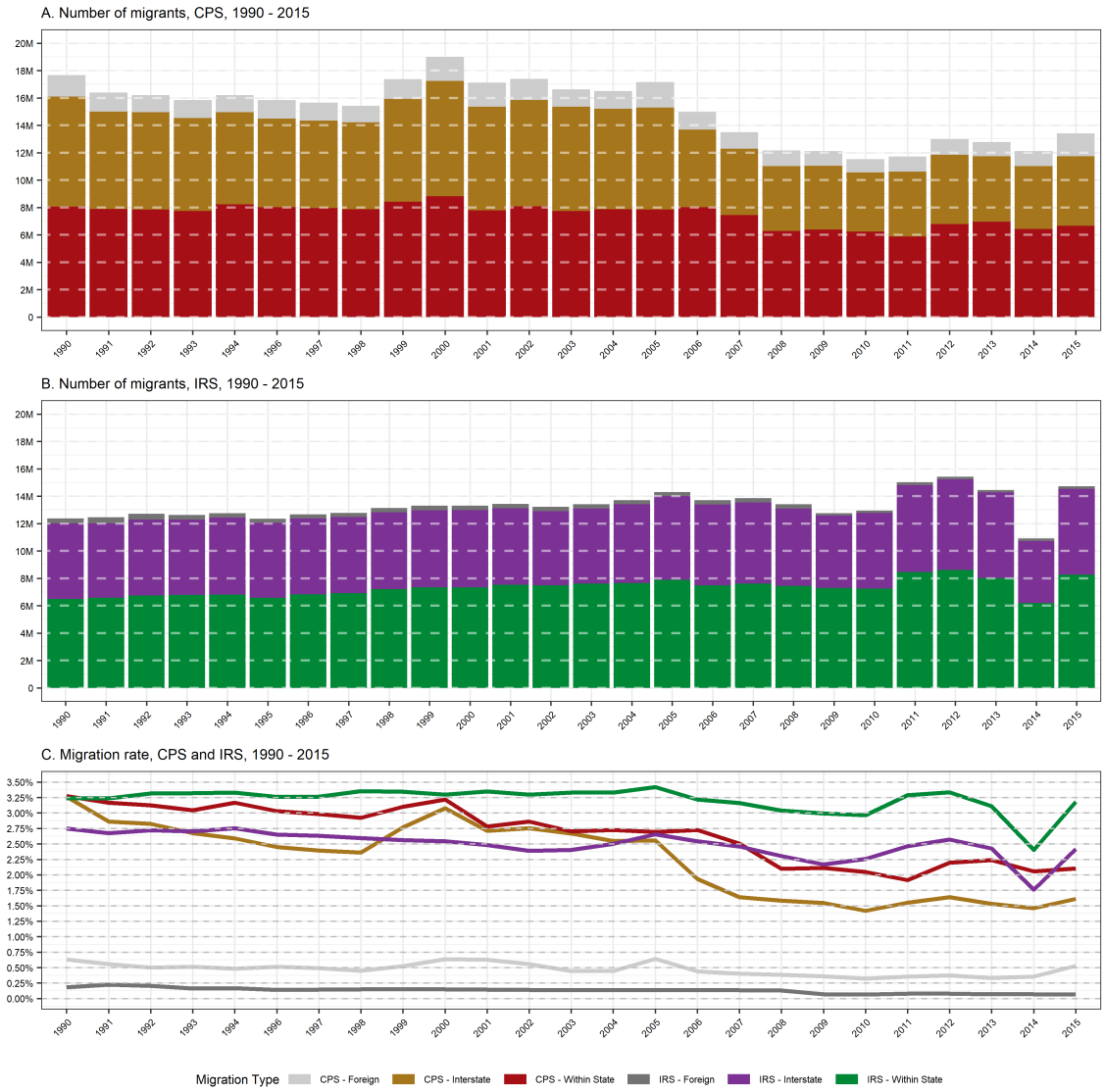


Figure 1.2: Within-state, interstate, and foreign migration counts and rates, CPS and IRS, 1990-2015

is that the suppression and censoring found in publicly available data is not present in the restricted access version. In addition, academic and government researchers are able to combine different restricted-access federal data sources to study migration (Foster et al. 2018). Finally, Bell et al. (2015) examine internal migration intensities across a number of countries and find that producing comparable estimates is difficult because cross-national internal migration data are not directly comparable.

Chapter two of this dissertation engages with and enhances migration data from the Internal Revenue Service to investigate the spatiality of internal migration during the 1990 through 2015 period. These data are used throughout this dissertation for four specific reasons: the data's availability (free!), the trusted source offering the data's provenance (the IRS), the temporal coverage (1990 through 2015), and the scale at which the data are presented (US counties). The research presented in this dissertation investigates the dynamics of how households move by making extensive use of the IRS's county-to-county household migration data. The IRS's county-to-county household migration data feature migration between counties in the US. A strength of the IRS data is yearly measurement of movement and the near-population scale at which the data are measured. As these data are publicly available and originate from different decades of information processing and storage, a fair amount of work must be carried out in order to harmonize and extend the usefulness of these data. Extending the usefulness of county-to-county household migration data enables the subsequent analyses featured in chapters three and four. Accordingly, the research goals of this dissertation are to extend a publicly available dataset, illustrate the spatiality of household migration preferences over time, and show how household mobility is conditioned by place-based characteristics and the metropolitan classifications of the origin and destination counties.

In chapter two I describe the technique I used to enhance the publicly available IRS county-to-county household migration data. The as-downloaded county-to-county household migration data feature the number of tax-returns - a proxy for the number of households - and the number of exemptions - a proxy for the number of individuals - residing in and mov-

ing between counties. The aggregate adjusted gross income moving between counties is also reported for years 1992 through 2015. The 26 years' worth of data feature several different schemas in several different formats. The spatial unit to which households and individuals are aggregated to, US counties, enables an analysis of household movement within-states and across state lines. As downloaded, the data are not readily useable without a great deal of examination and making comparisons across years is difficult. I detail the process of identifying four different epochs of IRS county-to-county household migration data and the development of an information schema flexible enough to accommodate the subtle, yet important, differences in each epoch to harmonize 26 years' worth of county-to-county household migration data.

The application of the consistent information schema enabled the reconciliation of unbalanced and missing flows. Because the IRS county-to-county data are offered in both a format of households entering a county (incoming migration) and a format of households leaving a county (outgoing migration), when properly arranged, each flow should appear twice, and the magnitude of the flows should be the same. This bookkeeping exercise revealed discrepancies in the data and were reconciled to produce a consistent set of household flows (returns) and flows of people (exemptions). My work harmonizes and reconciles the flows of households, people and money between counties for years 1990 through 2015. In total, I harmonized 5.2 million records.

Other researchers have also engaged with the IRS county-to-county data. Notably, Hauer and Byars (2019) implement a data harmonization scheme for the IRS county-to-county migration data for data pertaining to 1990 through 2010. Hauer and Byars harmonize tax exemptions only while I harmonize both returns and exemptions and estimate the origins and destinations of the censored returns leading to an additional 500K county-to-county records per year representing an additional 1.6 million households. Comparing the results from my harmonization process from 1990 through 2010 to Hauer and Byars's work, I identified several hundred additional county-to-county flows through the establishment of a flexible and detailed information schema. In this regard, this dissertation is about migration data

and what the data represent: migration. These simultaneous perspectives enable a more thorough investigation of the spatiality of internal migration.

Like most data disseminated from government agencies, the IRS data feature a level of suppression to protect tax payer's privacy and confidentiality (Gross 2009). The mechanism in place protecting confidentiality and privacy is to suppress all county-to-county flows less than 10 during the 1990 through 2012 period and suppress all flows less than 20 during the 2013 through 2015 period. The suppressed flows for each county are included in an aggregate remainder within-state migration total or an aggregate remainder interstate migration total. An aggregate remainder within-state migration total is the number of households that moved from a focal county to another county within the same state of origin. In this regard, a fully reported county-to-county flow is when there is a known county origin and known county destination. A censored county-to-county flow is when there is a known origin, but not a known destination, or a known destination, or a known origin. The interstate migration aggregate remainder totals feature the outcome of the censoring. In any given year, a little over 50-percent of all county-to-county household movement is fully reported within-state movement and approximately 25-percent of all county-to-county household movement is fully reported interstate movement. Combined, in any given year, 75-percent of all county-to-county household movement feature a fully reported origin and destination. The outcome of this suppression is that in any given year, about 5-percent of county-to-county movement is included in the aggregate remainder within-state migration sum and 20-percent of county-to-county household movement is included in the aggregate remainder interstate migration sum. I developed an algorithm to estimate the origins and destinations of the remaining 25-percent of county-to-county household movement.

There is a precedent for estimating censored data from the US government. The County Business Pattern data feature employment numbers by industry by county. Employment numbers by industry are suppressed when the number of firms in a county or the number of employees in an industry or firm does not clear a threshold. Isserman and Westervelt (2006) estimate the suppressed employment numbers in the US Census Bureau's 2002

County Business Pattern data using a linear optimization programming technique. The authors estimate the suppressed number by taking advantage of the hierarchical nature of the County Business Pattern data: reported county employment, known national employment, and known employment by industry totals. Zhang and Guldmann (2015) estimate the suppressed employment number in the 1999 through 2006 County Business Pattern data by taking advantage of the yearly trends in employment. Estimating the origins and destination of the county-to-county household movement data is important because it removes any potential biases resulting from using only 75-percent of the county-to-county data. This in turn enables a more robust examination of intrastate and interstate movement by looking at household movement to and from large population centers and household movement to and from smaller population centers.

The algorithm I developed uses a combination of regression modelling and linear optimization programming to estimate the origins and destinations of the suppressed county-to-county flows. The regression modelling components generate the upper bounds of a flow between any two counties without reported household movement. The linear optimization programming component determines the final value of the suppressed county-to-county flow by satisfying approximately 9.8 million constraints. The constraints ensure that the estimated values do not exceed an upper bound and the sum of all outgoing households matches the sum of all incoming households for all counties. The estimated origins and destinations for the suppressed county-to-county flows is accomplished by estimating the value of a suppressed county-to-county flow: estimating a county-to-county flow estimates an origin or a destination. Implementing this algorithm for each year of county-to-county movement data triples the intrastate migration complete case count and increases the interstate migration complete case count by 13-fold. This increased record count enables a more robust investigation into the spatiality of county-to-county household migration. For example, the enhanced data show that approximately 35-percent of moves in every year of the study period are to a directly adjacent county while the as-available county-to-county migration data show that 50-percent of all moves are to a directly adjacent county.

1.5 Modelling household destination preferences

Both chapter three and chapter four of this dissertation make use of the enhanced county-to-county migration data. In chapter three, I fit production-constrained, origin-specific spatial interaction models for approximately 95-percent of counties in each year in the 1990 through 2015 period. A production-constrained, origin-specific spatial interaction model considers outgoing migration only and an origin-specific model produces parameter estimates specific to each origin. Therefore, there are a little over 3K models produced for each year: one model for each county in each year for a total of 77K models. These models are estimated using a generalized linear modelling framework. During the 1990 to 2015 period, each county in the reported county-to-county household migration data sends households to approximately 25 destinations (and receives households from 25 origins), on average. The interquartile range covers 8 households to 18 households, on account of fewer, more populous counties sending households to more destinations. While the average number of destinations for any given county is 25, slightly less than 600 counties send households to at least 25 destinations; about 20-percent of all counties in the US. Using the 10 observations-per-explanatory- variable heuristic in a regression modelling framework Babyak (2004), Jenkins and Quintana-Ascencio (2020), and Troutt (2006), this limits the number of counties for which a robust model can be fit. After estimating the complete origins and destinations of the county-to-county migration data, 95-percent of counties feature at least 25 destinations and the average number of destinations per county is 212. Not only does the enhanced county-to-county migration data feature a greater number of counties for which a model can be estimated, the number of destinations per county increased. From a modelling perspective, this reduces potential biases originating from omitted observations.

I used the enhanced county-to-county migration data to fit approximately 3K production-constrained, origin specific, spatial interaction models per year for a total of 77K spatial interaction models. Each spatial interaction model featured the same four explanatory variables: the number of households in the destination, a value indicating a county's relative

spatial location in relation to other counties, the distance in miles between the origin and the destination, and a variable indicating if the movement is to a within-state or an out-of-state destination. The results of the models indicate the destination preferences of households. Even though the value of the preference is relative to the number of migrants each county sends, the values are comparable across space and over time. Visualizing the results of the spatial interaction models illustrates a spatiality to the preferences of household migration. Often there is an East/West distinction, but there are also distinct spatialities in different census divisions and regions. During the 26-year period, counties in the west and the rust belt were drawn to destinations with relatively less people when compared to the origin population. The preference for relatively more populated areas increased over time for counties in the south, the northeast, and in the plains. The spatiality of the preference for accessible destinations is most distinct and most consistent over time. Counties in western states move to relatively more accessible areas while counties in eastern states move to relatively less accessible areas. This finding is in part due to the spatial configuration of counties in the US: counties in the eastern states are smaller in area than counties in western states. Over time, households exhibited growing preferences in shorter distance moves. Especially so in counties in the east and south. Counties in California, Florida, Ohio, North Carolina, and upstate New York exhibit strong preferences for out-of-state destinations while counties in Montana, Idaho, Wyoming, Nevada, Utah show strong preference for in-state destinations.

1.6 Explaining moves to, from, and between metropolitan and non-metropolitan counties

The fourth chapter in this dissertation also makes use of the enhanced county-to-county migration data to showcase the effect of origin and destination place-based characteristics on household movement between and to metropolitan and non-metropolitan counties. By combining the reported county-to-county movement with the estimated county-to-county movement, all households moving into a county and all households moving out of a county are known. By determining the metropolitan status of the origin county and the destination

county, I can determine the proportion of moves between metropolitan counties, between non-metropolitan counties, and moves between metropolitan and non-metropolitan counties. In each year of the study period, approximately 65-percent of household moves are from one metropolitan county to a different metropolitan county. About 12-percent of household moves are from non-metropolitan counties to metropolitan counties and about 12-percent are from metropolitan counties to non-metropolitan counties. The remaining approximately 11-percent of moves are moves between non-metropolitan counties.

A feature of the publicly available county-to-county household migration data is that the age of the household (and other demographic and socio-economic characteristics) is not known. Only the number of households moving from one county to another is known. While it is not possible to tabulate the rates of migration by age between counties, I can look at the degree to which specific age groups propel and attract households between different types of counties using a spatial interaction model. It is well documented that there are differentials in migration rates by age group with mobility rates peaking for young adults and a smaller peak seen in retirement age adults (Pandit 1997; Rogers et al. 2002). Young adults are drawn to urban areas and central cities for school and employment while those of retirement age move to more rural. In this light, migration is a function of age and space. These differentials are reflected in the propulsive and attractive potentials of different age groups. I show how the 30-39 age group propels households between metropolitan counties and how the 50-59 age group attracts households to non-metropolitan counties. In total, I show how each nine different age groups contribute to household movement by the four different types of movement over a 26-year period. I also include economic indicators and outdoor amenity indicators. I show how households move to lower cost of living areas and that high amenity non-metropolitan counties draw additional households.

The three empirical chapters in this dissertation examine internal household migration in the US during the 1990 through 2015 period. Through curation, harmonization, and enhancement of a publicly available migration dataset, I investigate the spatiality of migrating household preferences and how place-based characteristics influence mobility. The next

three chapters describe the data enhancement process, the spatial patterns in the preferences of the migrating households, and the determinants of metropolitan and non-metropolitan movement. The fifth and final chapter of this dissertation features a summary of the results, recommendations for future study, and observations on writing and internal migration during the COVID-19 global pandemic.

Chapter 2

ENHANCING IRS COUNTY-TO-COUNTY HOUSEHOLD MIGRATION DATA

Chapter two describes the technique I used to enhance publicly available internal county-to-county household migration data by estimating the origins and destinations of county-to-county flows that were previously censored for privacy reasons. Household flows between US counties are suppressed when the number of flows fails to meet a specific disclosure threshold. In any given year in the 1990 through 2015 period, approximately 25-percent of county-to-county flows are censored, meaning that they are missing spatial detail. Using only 75-percent of the data introduces biases in subsequent analyses. This paper describes the development and application of a data cleaning and enhancement algorithm used to estimate the censored flows. This algorithm is applied to the Internal Revenue Service's county-to-county household migration data for each year in the 1990 through 2015 period. The algorithm described in this chapter results in an approximately 3-fold increase in within-state county-to-county household migration records and a 13-fold increase in interstate county-to-county household migration records in any given year in the period. Combined, an additional 1.6 million households per year can be included in subsequent analyses.

The first of this chapter's five substantive sections features a description of the data and the data schema I used to harmonize the 26-years' worth of data. The second section features yearly trends present in the migration data and presents some of the challenges in working with the county-to-county migration data. The third section describes the types of missing data present in the publicly available data and the fourth section describes the algorithm I used to estimate the complete set of origins and destinations for the censored values: a combination of count-data modeling and linear optimization programming techniques.

The final section features a comparison of the reported returns and a combination of the reported returns and the estimated remainder returns. There are two outcomes from the work undertaken in this chapter. The first is a set of harmonized, county-to-county household movement data with a schema consistent across the entire 26-year period enabling more consistent and immediate analyses. The second is that by estimating the complete origins and destinations of an additional 25-percent of migration data for each year, the average number of destinations per county increased from 25 to 212 (an 8x increase) enabling more robust, more thorough, and less-biased analyses of internal migration in the United States.

2.1 IRS county-to-county household migration data

The county-to-county migration data are produced by the IRS'S Statistics of Income branch and are "based on year-to-year address changes reported on individual income tax returns filed with the IRS" (*SOI Tax Stats - Migration Data — Internal Revenue Service* 2018). The data do not track people who are not required to file (low income and the elderly) nor people who file more exemptions, usually more wealthy people (Gross 2009). In effect, these data track households that are paying taxes and participate in the labor force to some capacity. These data are accessible over the World Wide Web and available free of charge. At the beginning of this writing in 2019, the IRS made available county-to-county migration data for tax filing years 1991 through 2016 (a tax filing year is the year in which taxes are filed for the previous year). As of October of 2021, migration data are now available through tax filing year 2019. There are minimal metadata associated with the county-to-county household migration data and what metadata the IRS does provide describes the contents of the data and not necessarily how the data are presented to the end user (Gross 2009). For clarity, the year associated with any given year of data is based on the previous year and not the tax filing year. For example, movement data for 1999 uses the 1999-2000 data. The move is assumed to have been made in 1999 and the tax return was filed in 2000.

In general, the data as downloaded from the IRS feature the number of tax returns (a proxy for households), the number of exemptions (a proxy for people), and the aggregate

gross income. These data are available in both an inflow (the number of tax returns coming into a county) and outflow (the number of tax returns leaving a county) format. Overall, the data for each year in the 26-year period feature the same variables with minor variance: state destination, county destination, state origin, county origin, state abbreviation, county name, number of tax returns, number of exemptions, and the aggregated gross income. The aggregate gross income is not available in years 1990 and 1991 but is available for all years beginning in 1992. In addition to the specific county-to-county movement, there are aggregate total records used to indicate movement of households between larger geographic scales. For example, movement into a county from other counties within the same state of origin or same census region of origin. Finally, there is a record indicating the number of returns filed by households that did not move. Households that move within county are included the non-mover category.

Consistent across all tax-filing years is that records are not immediately identifiable upon initial download from the IRS. In other words, each record, without extensive examination, is not readily distinguishable from a county-to-county movement or an aggregate remainder or a non-mover record. There is no single column available to consult that informs the end user of the type of record. It is important to note that while all movement between counties is recorded, not all movement is recorded the same way. For migration from one county to another featuring a movement of less than 10 households (returns) for years 1990 through 2012 and a movement of less than 20 households for years 2013-2015, a flow is added to a total featuring a more coarse scalar aggregation. This aggregation is done to protect taxpayer's confidentiality, privacy, and disclosure risk. For example, if the flow of households into a focal county is less than 10, but within the same state of origin, the flow is added to the within-state aggregate line. Similar operations are performed if the flow is from a different state or from a different census region. Unique to these data is how migration is presented in multi-scalar aspects. These data capture household movement across the US for each year in a 26-year period and those movements are gathered in aggregate and reported at the county scale, when feasible.

There is precedent for working with migration data from the IRS. Isserman et al. (1982) examine the usefulness of the IRS migration data and other federal data sources. Plane (1999) uses the IRS migration data to track income change and Plane et al. (2005) use the IRS migration data to examine migration to, from, and between larger and smaller metropolitan areas. This chapter is focused on enhancing IRS migration data in order to more thoroughly investigate internal migration in the US and not about internal migration. In addition, other researchers have also harmonized these data (Hauer and Byars 2019)[1].

There is also a precedent for estimating censored data from the US government. The County Business Pattern data feature employment by industry by county and employment numbers by industry or by county are suppressed when the number of firms in a county or the number of employees in an industry or firm does not clear a threshold. Isserman and Westervelt (2006) estimate the suppressed employment numbers in the US Census Bureau's 2002 County Business Pattern data using a linear optimization technique. Zhang and Guldmann (2015) estimate suppressed employment numbers in the 1999 through 2006 County Business Pattern data by taking advantage of yearly trends in employment numbers. Before any kind of analysis can be undertaken, the IRS county-to-county household migration data must first be placed into a consistent schema. The first step in harmonizing these data is identifying distinct epochs based on similarities and differences within the as-downloaded data.

2.2 Data epochs

As initially acquired, the IRS county-to-county household migration data, are not in a standard format nor in a consistent format[2]. Subsequent analyses and modelling are simply not

[1]I began working with this dataset in 2013. I developed version 1.0 of the harmonization technique in 2015. In 2020 I completed the version of the harmonization technique featured in this chapter. I first learned of Hauer and Byars's work in April of 2021.

[2]The IRS's Statistics of Income branch disseminates these data. While what the data represent is similar across time, its presentation is not. It is not too hard to imagine different generations of IRS analysts preparing these data with different schema mandates and ideas and available technology. For example, a user id of "blmccl00" is listed as the author of the 1993 Excel files and a user id of "raschw00" is listed as

possible without harmonizing these data. Four distinct data epochs were identified based on a consideration of three components of the data: 1) the uncompressed file format of the data; 2) the structure of the data (fixed width versus comma delimited); and 3) the values used to indicate censoring, aggregation, and intermediate totals within the data. Overall, the older county-to-county migration data are more challenging than the newer county-to-county migration data. Python 3.7 (Van Rossum and Drake Jr 1995) was used to process each year of data and parse each record into various components and fields. Custom functions making use of the Pandas library for python (McKinney 2010) were written to import and format the data. A key contribution of this research is the preparation, error checking, and modernization of these data. The end product of processing these data is the placement of these data into a SQLite (Hipp 2020) relational database.

2.2.1 Epoch 1: 1990 and 1991

The data for the years in Epoch 1 are available on a state-by-state basis for both the incoming and outgoing flows. That is, there is an incoming migration file and an outgoing migration file for the 50 states, Washington DC, and an incoming migration file and an outgoing migration file for foreign locations. In total, there are 104 files for 1990 and 104 files for 1991. Unique to years 1990 and 1991, the data are delivered in a fixed-width, off-set, and hierarchical text format (*.txt). Figure 2.1, page 23 features records for in-migration in 1990 for Bristol County, Rhode Island. There are multiple pieces of information in Figure 2.1 that have been identified by rows and columns for the purpose of discussion.

Row A features the reference county with the total number of movers into Bristol County, Rhode Island: 1,188 tax returns (households) and 2,184 exemptions (people). The Federal Information Processing Standard (FIPS) Code for Bristol County, Rhode Island is 44001. Rows in group B feature the counties of origin for movement into Bristol County with the attendant FIPS Codes for each of the origin counties, the returns, and the exemptions. Rows

---

the author the 2003 Excel files.

		1	2	3	4	5
A	44 001 Bristol Total Migration	RI	1188		2184	
B	44 007 Providence	RI	381	32.07	667	30.54
	25 005 Bristol	Ma	120	10.10	208	9.52
	44 005 Newport	RI	94	7.91	192	8.79
	44 009 Washington	RI	41	3.45	71	3.25
	44 003 Kent	RI	38	3.20	86	3.94
	25 017 Middlesex	Ma	33	2.78	50	2.29
	25 025 Suffolk	Ma	13	1.09	13	.60
	25 021 Norfolk	Ma	13	1.09	25	1.14
	09 001 Fairfield	Ct	12	1.01	18	.82
	12 099 Palm Beach	Fl	12	1.01	20	.92
	57 001 Foreign / Overseas	FR	12	1.01	16	.73
	36 061 New York	NY	11	.93	14	.64
	12 103 Pinellas	Fl	11	.93	21	.96
	25 023 Plymouth	Ma	10	.84	27	1.24
C	Region 1: Northeast		128	10.77	252	11.54
	Region 2: Midwest		32	2.69	75	3.43
	Region 3: South		163	13.72	312	14.29
	Region 4: West		64	5.39	117	5.36
D	44 001 County Non-Migrants		17777		39191	

Figure 2.1: Data format for Epoch 1: 1990-1991

in group C feature movement, by Census Region, into Bristol County. Row D features the total number of returns (households) that did not move (17,777) and the total number of exemptions (people) that did not move (39,191) in Bristol County. Column 1 features the state abbreviations for the reference county and origin counties. Column 2 features the number of households (tax returns) flowing into Bristol County and column 3 expresses those values a percentage. Column 4 showcases the number of people (exemptions) that moved into Bristol County and column 5 expresses those values as a percentage of the total. For the data pertaining to 1990 and 1991, the reference county features no indention while subsequent rows feature indention indicating movement or the lack thereof. Field widths were identified based on the starting position of various characters.

2.2.2 Epoch 2: 1992 through 1994

The data in Epoch 2, 1992 through 1994, feature in-migration and out-migration files for each of the 50 states, Washington DC, and foreign movement. In total there are 312 files for this period. Unlike the 1990 and 1991 years, the data are delivered in a Microsoft Excel format. Figure 2.2, page 25, features a selection from the 1993 out migration file from South Carolina. In this format, rows five through eight of columns A through I delineate the field names while rows nine through 20 depict migration out of Abbeville County, South Carolina. Row nine features information about the reference county, rows 10 through 16 showcase destinations of households migrating from Abbeville County. Rows 17 through 19 feature data aggregated to different scales. Again, this is done to protect taxpayer confidentiality and privacy and reduce disclosure risk. Finally, row 20 features the counts of households that did not migrate. The aggregate money associated with a flow or non-movers (column I) is first introduced in 1992. The aggregate income value is the gross sum of aggregate household income in thousands of dollars. In this case, the flow of households from Abbeville County, South Carolina in 1993 exited the county with a combined gross income of approximately $8.3 million (1993) dollars.

Similar to the data for 1990 and 1991, there is no way to readily select records that pertain to county-to-county movement (row 10, for example) or records that pertain to an aggregated total (row 17, for example) based on a single value. Upon inspection of the data, I learned that FIPS codes were applied to the aggregate rows as well. In this case, within-state migration has a FIPS code of 63020, regardless of the state in question. The FIPS code for within-state migration changes in subsequent epochs.

2.2.3 Epoch 3: 1995 through 2003

The files for data in Epoch 3, years 1995 through 2003, are presented in an incoming file and an outgoing file for each of the 50 states, Washington DC, and for foreign origins and destinations. Except for the 2003 data, there are 52 files for each year; the 2003 data do

	A	B	C	D	E	F	G	H	I
1	1993 - 1994 County to County Migration Outflow								
2	(Aggregate money amounts are in thousands of dollars)								
3									
4									
5	Migration from SouthCarolina		Migration into				Number of returns	Number of exemptions	Aggregate total money income
6									
7	State	County	State	County	State	State totals, county totals, and county by county detail			
8	FIPS Code								
9	45	001	00	001	SC	Abbeville (Total Migrant)	466	964	8,316
10	45	001	45	047	SC	Greenwood	112	226	2,177
11	45	001	45	007	SC	Anderson	104	212	1,792
12	45	001	45	045	SC	Greenville	29	66	677
13	45	001	45	059	SC	Laurens	17	38	230
14	45	001	45	065	SC	McCormick	16	38	376
15	45	001	45	083	SC	Spartanburg	11	21	158
16	45	001	45	079	SC	Richland	10	15	119
17	45	001	63	020	XX	Same State	50	92	915
18	45	001	63	021	XX	Same Region, Diff. State	94	206	1,632
19	45	001	63	022	XX	Different Region	23	50	240
20	45	001	63	050	SC	County Non-Migrant	7,779	17,614	189,665

Figure 2.2: Data format for Epoch 2: 1992-1994

not have the foreign movement files. There are 934 files in total for Epoch 3. The files as downloaded from the IRS were in a Microsoft Excel format. Figure 2.3 on page 26 features a subset of records from the 1999 data for Oregon for inflow migration. There are several changes to this record format. First, there are additional rows featuring aggregate totals for the entire state of Oregon (rows 9 through 13) and for Baker County, Oregon (rows 14 through 18). Rows 19 through 29 feature the county origins of households moving into Baker County. Finally, rows 30 and 31 feature aggregates of exemptions for when the minimum disclosure threshold was not met. In this case, there were 65 returns (households) from other counties in Oregon and 172 returns from other states that migrated to Baker County, Oregon in 1999. A substantial change introduced with this epoch of data is that the aggregate flows were assigned a different FIPS code. In Epoch 1, flows from other parts of the state were not assigned a FIPS code, these records were merely indented four spaces. In Epoch 2, flows from other parts of the state were assigned a FIPS code of 63020. In Epoch 3, flows from other parts of the state were coded as 58000.

	A	B	C	D	E	F	G	H	I
1	1999 - 2000 County to County Migration Inflow								
2	(Aggregate money amounts are in thousands of dollars)								
3									
4									
5	Migration into Oregon		Migration from				Number of returns	Number of exemptions	Aggregate adjusted gross income
6									
7	State	County	State	County	State	State totals, county totals,			
8	FIPS Code					and county by county detail			
9	41	000	96	000	OR	Total Mig - US & For	106,648	193,862	3,941,071
10	41	000	97	000	OR	Total Mig - US	105,094	191,557	3,904,301
11	41	000	97	001	OR	Total Mig - US Same St	57,027	103,209	2,025,582
12	41	000	97	003	OR	Total Mig - US Diff St	48,067	88,348	1,878,719
13	41	000	98	000	OR	Total Mig - Foreign	1,554	2,305	36,769
14	41	001	96	000	OR	Baker County Tot Mig-US & For	432	930	12,700
15	41	001	97	000	OR	Baker County Tot Mig-US	432	930	12,700
16	41	001	97	001	OR	Baker County Tot Mig-Same St	235	496	6,782
17	41	001	97	003	OR	Baker County Tot Mig-Diff St	197	434	5,918
18	41	001	41	001	OR	Baker County Non-Migrants	5,096	11,558	167,564
19	41	001	41	061	OR	Union County	32	68	848
20	41	001	16	001	ID	Ada County	25	59	764
21	41	001	41	059	OR	Umatilla County	24	48	846
22	41	001	41	051	OR	Multnomah County	22	41	691
23	41	001	41	017	OR	Deschutes County	21	44	465
24	41	001	41	005	OR	Clackamas County	14	22	596
25	41	001	41	023	OR	Grant County	13	36	315
26	41	001	41	045	OR	Malheur County	12	32	281
27	41	001	41	067	OR	Washington County	12	23	421
28	41	001	41	039	OR	Lane County	10	17	325
29	41	001	41	047	OR	Marion County	10	22	245
30	41	001	58	000	SS	Other Flows - Same State	65	143	1,749
31	41	001	59	000	DS	Other Flows - Diff State	172	375	5,154

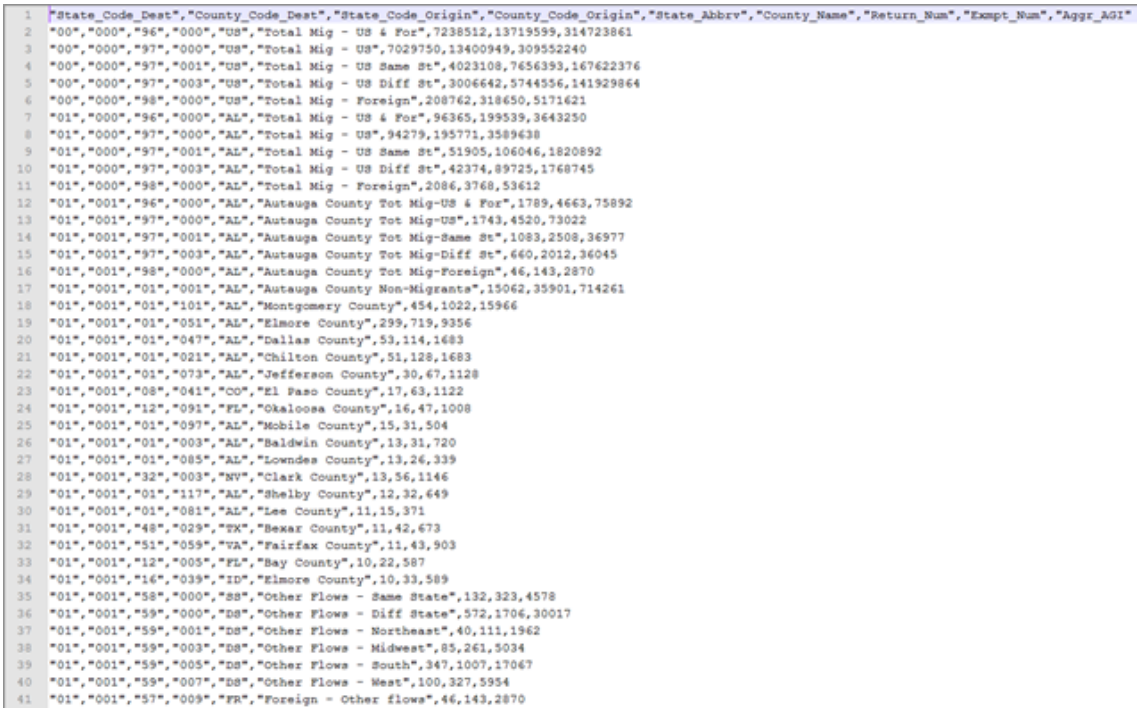
Figure 2.3: Data format for Epoch 2: 1995-2003

2.2.4 Epoch 4: 2004 through 2015

The final and current epoch of data begins in 2004 and extends through the most recent release of data featured in this chapter, 2015. There are several changes to this epoch of data. First, the data are delivered in a comma separated value format (*.txt) instead of a Microsoft Excel format. Second, rather than individual files for each state, the data are delivered in one file each for incoming flows and outgoing flows. Accordingly, there are 24 files for this epoch of data. Figure 2.4 on page 28 features a selection of records from the 2005 county inflow. Like the previous epoch of data, there are aggregate totals for the US (rows 2 through 6) and for each state (rows 7 through 11 for Alabama in this case). Rows 12 through 17 feature aggregate totals for Autauga County, Alabama, and rows 18 through 34 feature the origins of movement for migration to Autauga County. Rows 35 through 41 feature aggregate remainders of migration to Autauga, County. Epoch 4 features the most consistent data formatting. However, the values of the aggregate remainder codes changed between Epoch 3 and Epoch 4.

In total, there are 739 files pertaining to incoming flows and 739 files pertaining to outgoing flows for a total of 1,478 files for the 26-year timespan covered by the data. Across the incoming files, there are approximately 2.79 million records pertaining to incoming flows and 2.78 million records pertaining to outgoing flows for a total of approximately 5.58 million records. Certainly not a lot of data, at the time of this writing, but there is a depth to these data that enables a new and multi-faceted approach to the understanding of internal migration in the United States.

Each of these data epochs feature internal migration data presented in a somewhat similar but not readily accessible fashion. Of note is that Epoch 4 features the most attribute-rich format: specific indicators via FIPS codes for each type of record indicating distinct record types: county-to-county movement, aggregate totals, and remainder totals. While Epoch 4 features something most closely resembling an accessible schema, it does not enable the rapid and correct identification of records for specific analysis. As initially received, records must



```
"State_Code_Dest","County_Code_Dest","State_Code_Origin","County_Code_Origin","State_Abbrv","County_Name","Return_Num","Exmpt_Num","Aggr_AGI"
"00","000","96","000","US","Total Mig - US & For",7238512,13719599,314723861
"00","000","97","000","US","Total Mig - US",7029750,13400949,309552240
"00","000","97","001","US","Total Mig - US Same St",4023108,7656393,167622376
"00","000","97","003","US","Total Mig - US Diff St",3006642,5744556,141929864
"00","000","98","000","US","Total Mig - Foreign",208762,318650,5171621
"01","000","96","000","AL","Total Mig - US & For",96365,199539,3643250
"01","000","97","000","AL","Total Mig - US",94279,195771,3589638
"01","000","97","001","AL","Total Mig - US Same St",51905,106046,1820892
"01","000","97","003","AL","Total Mig - US Diff St",42374,89725,1768745
"01","000","98","000","AL","Total Mig - Foreign",2086,3768,53612
"01","001","96","000","AL","Autauga County Tot Mig-US & For",1789,4663,75892
"01","001","97","000","AL","Autauga County Tot Mig-US",1743,4520,73022
"01","001","97","001","AL","Autauga County Tot Mig-Same St",1083,2508,36977
"01","001","97","003","AL","Autauga County Tot Mig-Diff St",660,2012,36045
"01","001","98","000","AL","Autauga County Tot Mig-Foreign",46,143,2870
"01","001","01","001","AL","Autauga County Non-Migrants",15062,35901,714261
"01","001","01","101","AL","Montgomery County",454,1022,15966
"01","001","01","051","AL","Elmore County",299,719,9356
"01","001","01","047","AL","Dallas County",53,114,1683
"01","001","01","021","AL","Chilton County",51,128,1683
"01","001","01","073","AL","Jefferson County",30,67,1128
"01","001","08","041","CO","El Paso County",17,63,1122
"01","001","12","091","FL","Okaloosa County",16,47,1008
"01","001","01","097","AL","Mobile County",15,31,504
"01","001","01","003","AL","Baldwin County",13,31,720
"01","001","01","085","AL","Lowndes County",13,26,339
"01","001","32","003","NV","Clark County",13,56,1146
"01","001","01","117","AL","Shelby County",12,32,649
"01","001","01","081","AL","Lee County",11,15,371
"01","001","48","029","TX","Bexar County",11,42,673
"01","001","51","059","VA","Fairfax County",11,43,903
"01","001","12","005","FL","Bay County",10,22,587
"01","001","16","039","ID","Elmore County",10,33,589
"01","001","58","000","SS","Other Flows - Same State",132,323,4578
"01","001","59","000","DS","Other Flows - Diff State",572,1706,30017
"01","001","59","001","DS","Other Flows - Northeast",40,111,1962
"01","001","59","003","DS","Other Flows - Midwest",85,261,5034
"01","001","59","005","DS","Other Flows - South",347,1007,17067
"01","001","59","007","DS","Other Flows - West",100,327,5954
"01","001","57","009","FR","Foreign - Other flows",46,143,2870
```

Figure 2.4: Data format for Epoch 2: 2004-2015

be selected through complex and verbose structured query statements. It is not possible to select all county-to-county movement records based on querying a single database field; there is no unifying schema across the timespan of the data that permits this. It is to this concept that this section now turns: the instantiation of a normalized database schema permitting a ***quick, easy, accurate, and replicable*** selection of records for internal migration analysis.

2.3 Instantiation of a database schema

To review, all four of the data epochs feature the following columns organized by incoming flow or outgoing flow: origin/destination state FIPS code, origin/destination county FIPS code, destination/origin state FIPS code, destination/origin county FIPS code, origin/destination state abbreviation, origin/destination county name, the number of returns in the flow, the number of exemptions in the flow, and with the exception of the data for

Table 2.1: Counts of files and records by Epoch, 1990-2015

	Year	Incoming files	Outgoing files	Total files	Incoming records	Outgoing records	Total records
Epoch 1	1990	52	52	104	95,478	95,136	190,614
	1991	52	52	104	96,230	95,764	191,994
Epoch 2	1992	52	52	104	97,149	96,710	193,859
	1993	52	52	104	96,746	96,311	193,057
	1994	52	52	104	98,751	98,256	197,007
Epoch 3	1995	52	52	104	106,981	105,961	212,942
	1996	52	52	104	108,739	108,042	216,781
	1997	52	52	104	109,319	108,694	218,013
	1998	52	52	104	110,555	110,180	220,735
	1999	52	52	104	111,201	110,900	222,101
	2000	52	52	104	111,807	111,784	223,591
	2001	52	52	104	111,581	111,279	222,860
	2002	52	52	104	109,620	109,262	218,882
	2003	51	51	102	109,820	109,155	218,975
Epoch 4	2004	1	1	2	114,534	113,920	228,454
	2005	1	1	2	118,389	117,507	235,896
	2006	1	1	2	116,627	115,865	232,492
	2007	1	1	2	118,691	117,973	236,664
	2008	1	1	2	116,059	115,544	231,603
	2009	1	1	2	110,651	110,549	221,200
	2010	1	1	2	113,593	113,624	227,217
	2011	1	1	2	130,101	130,297	260,398
	2012	1	1	2	131,931	132,046	263,977
	2013	1	1	2	86,193	86,347	172,540
	2014	1	1	2	75,527	75,785	151,312
	2015	1	1	2	86,330	86,481	172,811
	Totals	**739**	**739**	**1,478**	**2,792,603**	**2,783,372**	**5,575,975**

years 1990 and 1991, the aggregate gross income in the flow. The data are presented in a similar but not identical format across epochs and different codes are used to indicate the type of aggregate remainder or aggregate total. I undertook the process of instantiating a database schema that classifies each record enabling both subsequent analysis and verifying the integrity and accuracy of the data.

Regardless of the data epoch, most records indicate county-to-county movement. The technique is to systematically identify all the records that are not county-to-county movement so that county-to-county records are the last records to be identified. Using basic text analysis and the regex library in Python 3.7, I generated a decision tree for classifying each record in each file. Each record was passed through the decision tree and a counter was initiated to track the number of times a record was classified. Multiple identifications were indicative of faulty or incomplete logic. Through an iterative process, each record was ultimately classified once and only once. The derivation of the epochs was based upon the differences in the logic used to classify each record in each year. Additional pre-processing was performed on the 1990 and 1991 data epoch to place the indented and fixed-width format into a columnar format. Starting in Epoch 3 (beginning in 1995) the aggregate interstate remainder row for different states were grouped by Census Region. For the data in Epoch 3, aggregate remainder rows for interstate migration were specified individually. If there were enough returns migrating to/from each census region, the other-state migration was disaggregated to each region. If there were not enough records, the migration was grouped to an aggregate interstate category, a same region category, or a different region category. As different ways of aggregating the data were introduced, repurposed, or removed, attention was paid to structuring the data. In total, I identified 27 different record types. Only one-third of the record types exist across all epochs of data.

An important and necessary aspect of the instantiation of this database schema is that it enables the discovery of inconsistencies and trends within the data. For example, in 1990 there were approximately three times as many returns from foreign origins when compared to 1991. The data in 2010 feature about 14K fewer records than in 2011. In addition, one

can see the year-to-year variation in migration flows and data changes. Vexing about this dataset is that some of the variation in the counts of migration flows is actual variation in migration and some of it is how the data were processed (DeWaard et al. 2020b; Pierce 2015, 2020). Finally, upon harmonization, the data were loaded into the above schema and stored in a SQLite database.

2.4 Yearly trends in internal household migration, 1990-2015

The first section of this chapter focused on the data as artifact: the structure of the files, the number of records, and the types of records. Using the harmonized data, I will discuss what these data represent: internal migration in the United States from 1990 through 2015 at different geographic scales and those trends over time. Beginning in 1990, 193K and 6.2M households migrated to and within the US for a total of 6.3M households on the move. Figure 2.5 on page 33 features the yearly counts, graphic A, rates, graphic B, and shares, graphic C, of incoming migration flows by foreign, interstate, and within-state classifications. The 6.3M figure translates to an internal migration rate of approximately 7.0-percent, with a dip during the Great Recession in the later 2000s and again in 2014. According to IRS documentation, processing changed in 2011 (DeWaard et al. 2020b; Pierce 2015). In addition, based on correspondence between the Missouri Census Data Center and Kevin Pierce, of the IRS's Statistics of Income Division, the IRS instituted new measures to combat identity theft which might have impacted the matching of returns (Pierce 2020). The count of households on the move reached a peak in 2012 with 8.1 million households migrating to (91K) and within (8.0M) the US while a peak migration rate was reached in 2005 with nearly 7.0-percent of returns being filed from migratory households.

The yearly rates of outgoing, internal migration are identical to the incoming rates of internal migration. Outgoing flows to foreign locations are, for the most part, less than incoming flows from foreign locations. From 1990 to 2008, between 2.5-percent to 3.5-percent of all US and foreign tax returns were filed by immigrating households. During this period, the US received more international migrants than it sent. In 2009, the number of

Table 2.2: Record types by Epoch

Record Type	Description	Epoch 1: 1990-1991	Epoch 2: 1992-1994	Epoch 3: 1995-2003	Epoch 4: 2004-2015
0	Total Migration: US and Foreign	Y	Y	Y	Y
1	Total Migration: US			Y	Y
2	Total Migration: Same State			Y	Y
3	Total Migration: Different State			Y	Y
4	Total Migration: Foreign			Y	Y
5	Non-migrants	Y	Y	Y	Y
6	Foreign: overseas	Y	Y	Y	Y
7	Foreign: Puerto Rico	Y	Y	Y	Y
8	Foreign: APO/FPO ZIPS	Y	Y	Y	Y
9	Foreign: Virgin Islands	Y	Y		Y
10	Foreign: Other flows	Y	Y	Y	Y
11	Other Flows: Same State	Y	Y	Y	Y
12	Other Flows: Same Region, Diff. State	Y	Y		
13	Other Flows: Different Region	Y	Y		
14	Other Flows: region 1 / northeast	Y	Y		
15	Other Flows: region 2 / midwest	Y	Y		
16	Other Flows: region 3 / south	Y	Y		
17	Other Flows: region 4 / west	Y	Y		
18	Other Flows: Different State			Y	Y
19	sub-aggregation: Different State - Northeast			Y	Y
20	sub-aggregation: Different State - Midwest			Y	Y
21	sub-aggregation: Different State - South			Y	Y
22	sub-aggregation: Different State - West			Y	Y
23	All migration flows: remainder	Y			
24	Suppress all flows		Y		
101	Same state county-to-county migration	Y	Y	Y	Y
102	Different state county-to-county migration	Y	Y	Y	Y

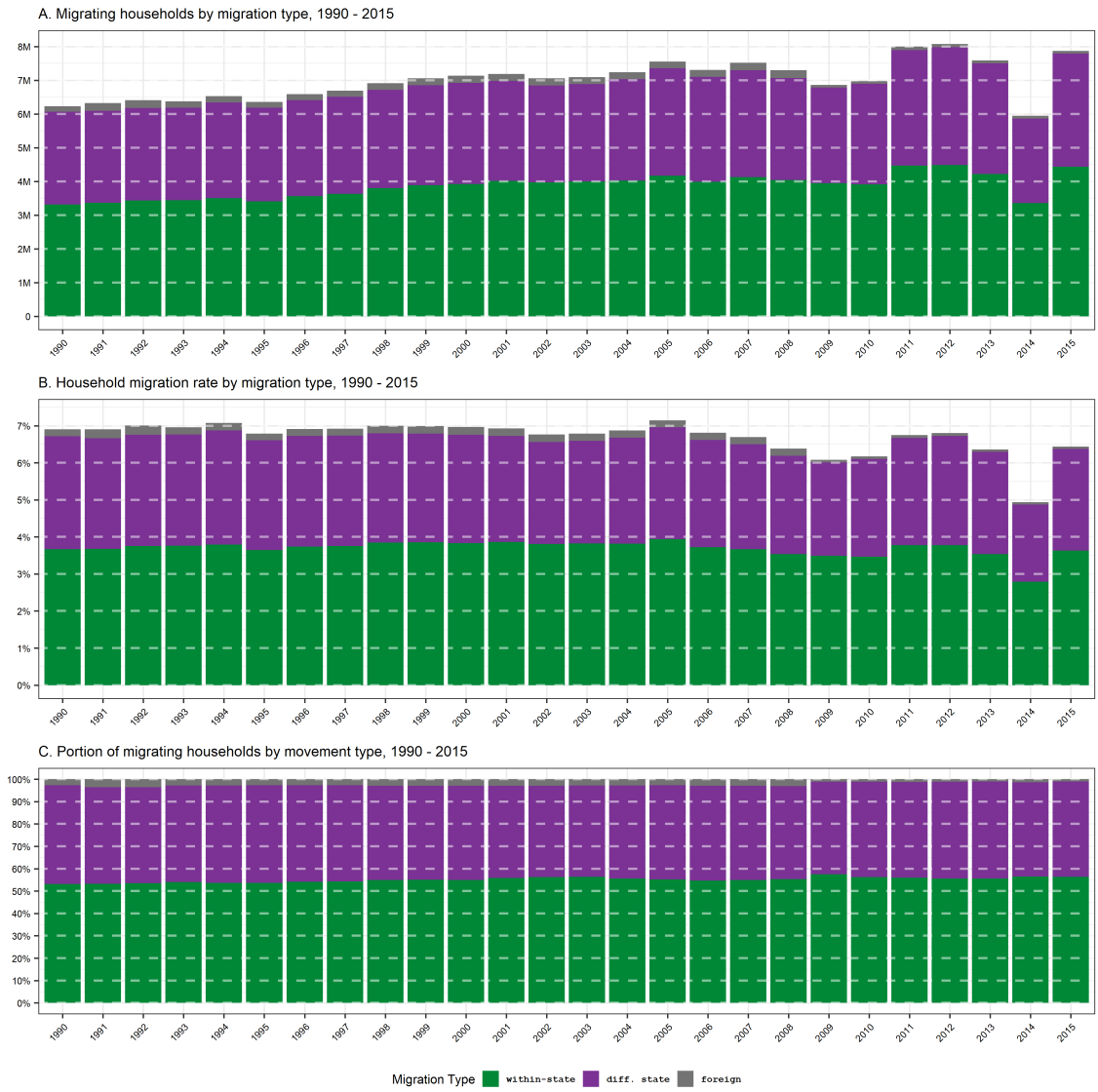


Figure 2.5: Yearly counts of household migration by origin classification, 1990-2015

immigrating households dropped sharply to less than 1.5-percent where it has remained ever since. Since 2009, the US sends more migrants than it receives.

In general, roughly six to seven percent of tax returns are filed by migratory households (except for 2014, in which the rate dropped to approximately five percent). Graphic C in Figure 2.5 depicts how the portions of households moving within a state, between states, and in from a foreign location have remained relatively steady across the 26-year time span. Of the households that move, approximately 50-percent of those households move to an area elsewhere within the same state of origin, slightly less than 46-percent move to a different state and upwards of 4-percent migrate from a foreign location. While these migration rates and counts reflect the movement of return, a proxy for households, the IRS data also include tallies of exemptions, a proxy for the number of people in the household. Trends in migration rates of exemptions is similar to trends in migration rates of returns.

These relatively simple graphs, given the specified categories, suggest that the internal migration rate in the US has experienced slight increases and decreases since 1990 with a noticeable dip during the Great Recession and the dip in 2014 due to processing changes. While foreign immigration and emigration have changed over time and rates of internal migration at the national scale have changed modestly over time, what are the trends in county-to-county movement over the 26-year period? Before I can answer that question, additional data analyses are necessary to fully understand and model the flows.

2.4.1 Aggregate remainder returns

While the foreign migration category is coarse (countries of origin or destination are not listed), the within-state and interstate migration categories can be disaggregated into the following: within-state reported, within-state remainder, interstate reported, and interstate remainder. The within-state and interstate reported categories indicate a known origin and known destination. For example, a certain number of households moved from Cook County, Illinois to Multnomah County, Oregon in a given year. The within-state and interstate remainder categories indicate that some number of households moved into a county from

elsewhere in the state or from a different state. For example, a certain number of households migrated to Multnomah County, Oregon from an unknown number of locations within Oregon and a certain number of households migrated to Multnomah County, Oregon from an unknown number of locations in states other than Oregon. Figure 2.6, page 36, showcases the counts, graphic A, and rates, graphic B, of migration by the disaggregated within-state and interstate categories. For each year, between 1.5 million and 2.2 million households are grouped in a remainder category. That is, for any given year, between 1.5 million and 2.2 million households only have a known destination or a known origin, but not both. Approximately 50-percent of migrating households are fully reported within-state, 5-percent of migrating households are grouped into a within-state remainder category, 25-percent of migrating households are fully reported interstate, and 20-percent of migrating households are grouped into an interstate remainder category. Even when accounting for increases in the number of households on the move and a decrease in mobility due to the Great Recession, the portions of households in each of these categories is remarkably consistent over time.

What kind of bias is introduced by not including the households in these aggregate remainder categories? What additional knowledge about internal migration is gained from estimating the origins and destinations of the suppressed flows and including the estimated flows in subsequent analyses? The prime motivation for estimating the missing flows is to eliminate potential bias introduced from working with only 75-percent of the observed county-to-county household flows. Subsequent sections in this chapter are motivated by the theory that flows between counties are a function of the population sizes of the origin and destination counties and the distances between the counties. There are therefore going to be more flows between larger counties and more proximate counties and fewer flows between smaller counties and more distant counties,*ceteris paribus*. If using only the reported flows, more populous counties will be overrepresented as flows between counties are only reported if there are at least ten households moving between any two counties. In other words, more populous counties are more likely to feature incoming or outgoing flows of ten households than less populous counties. The effect of this overrepresentation of more populous counties inflates

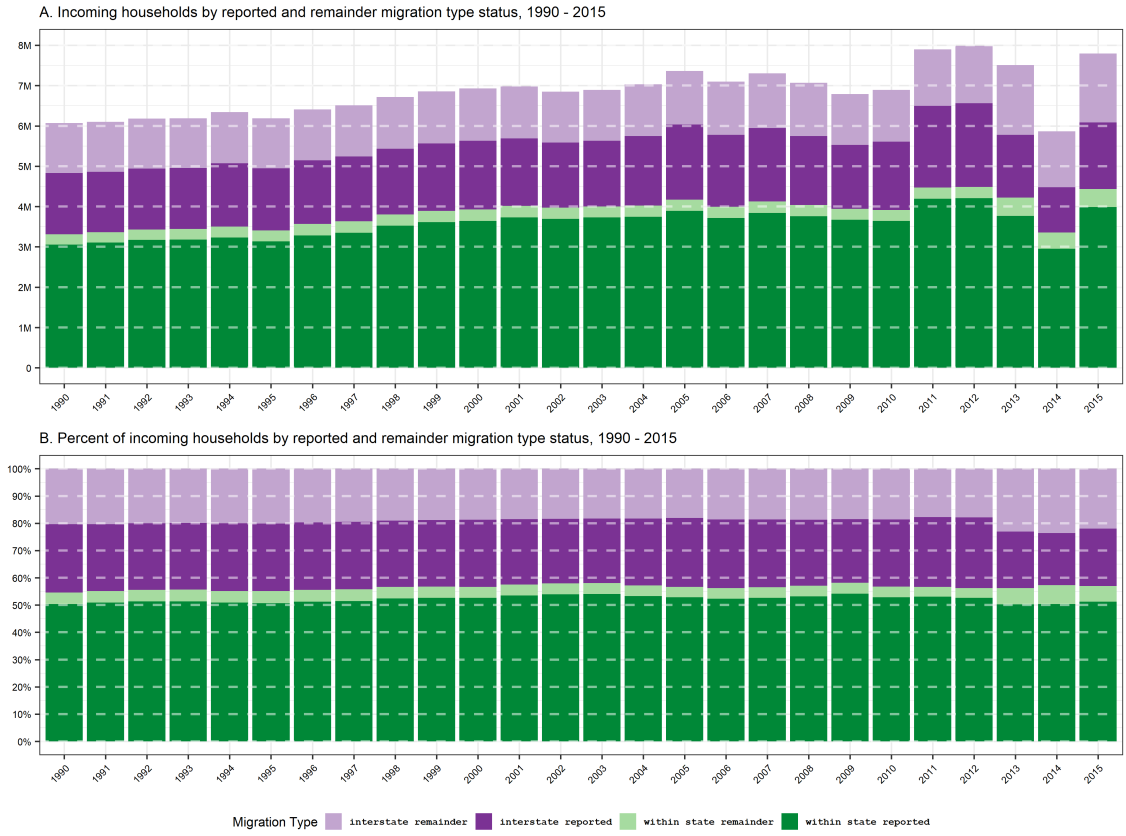


Figure 2.6: Counts and rates of incoming households, by reported and remainder migration categories, 1990-2015

the metropolitan migration rate. The next section features several hypothetical scenarios illustrating the effect of omitting and including additional county-to-county household flow records.

2.4.2 The utility of including aggregate remainder flows: a counterfactual example

Using the 1999 county-to-county household migration data, I will showcase the effect of including the estimated county-to-county flows through several counterfactual examples. These counterfactual examples showcase the changes in a distance decay coefficient from a production constrained, origin specific, spatial interaction model. The model specified for

each origin county i is:

$$T_{ij} = \alpha M_j + \delta A_j - \beta d_{ij} \tag{2.1}$$

Where T_{ij} is the flow of households from origin county i into destination county j. M_j is the mass term associated with destination county j, the number of non-movers at destination j. A_j is destination county j's accessibility score. The A_j term measures how distant each other county j's population is from county i. This term controls for spatial structure and has been shown to increase the performance of models of this type (Fotheringham 1983). Finally, d_{ij} is the geographic distance in miles in between origin county i and destination county j. The coefficients α, δ, β measure each county i's aggregate household preferences for large destinations, accessible destinations, and more proximate moves, respectively. The form of this model specification is based on work carried out by Yano et al. (2003) in which the authors examine the preferences of internal migrants in Japan and Britain. A more thorough discussion of the derivation of this modelling framework is included in the section titled 2.6, beginning on page 44.

These counterfactual scenarios focus specifically on the distance decay parameter, β. This parameter indicates the degree to which interaction decreases with distance. In most (if not all) migration systems, the coefficient is negative. The more negative the value, the more quickly interaction decreases with distance. Values near -1 indicate that interaction between two counties decreases linearly and values much greater or much less than -1 indicate that interaction between counties decreases exponentially. Because I am fitting a production constrained model for each county, each county will have its own distance decay coefficient. I am fitting these models on counties with at least 30 reported destinations to ensure reliable estimation of the coefficients.

Of the 3,143 counties in the US in 1999, 548 counties (about 1 in 6 counties) sent households to at least 30 unique destinations. Households from these 548 counties migrated to 2,648 counties (about 5 in 6). Of these 548 origin counties, the majority, 486, are metropolitan counties and the remaining 62 are non-metropolitan counties. The minimum reported

outgoing flow from these counties is 10, the average is 74, and the largest flow is 20,369 households[3]. In general, as county population size increases, so too does the number of destinations. To showcase the effect of including additional records, I will fit five models for each county. Each model will vary the number of additional records as follows:

- Model 1: The reported county-to-county migration values, known zero county-to-county migration flows, and setting all unobserved county-to-county migration values to zero.

- Model 2: The reported county-to-county migration values and the known zero county-to-county migration flows.

- Model 3: Only the reported county-to-county migration values.

- Model 4: The reported county-to-county migration values, the known zero county-to-county migration flows, and setting the unobserved county-to-county migrations cells to a random value between 0 and 9 until a total of each origin county's remainder outflow is reached.

- Model 5: The reported county-to-county migration values and setting the unobserved county-to-county migrations cells to a random value between 0 and 9 until a total of each origin county's remainder outflow is reached.

Models 4 and 5 use random data as a naïve proxy for the estimated county-to-county returns to showcase how the estimated returns will impact the distance decay coefficient. For the purposes of visualization and interpretation, counties are grouped into deciles according to the number of reported destinations. The destination count deciles are listed in the legend in Figure 2.7 on page 39. Grouping by decile is done to showcase the degree to which

[3]This 20.4K flow is the number of households that moved from Orange County, CA to Los Angeles County, CA. The second largest flow in 1999 was approximately 16K households moving from Los Angeles County, CA to Orange County, CA.

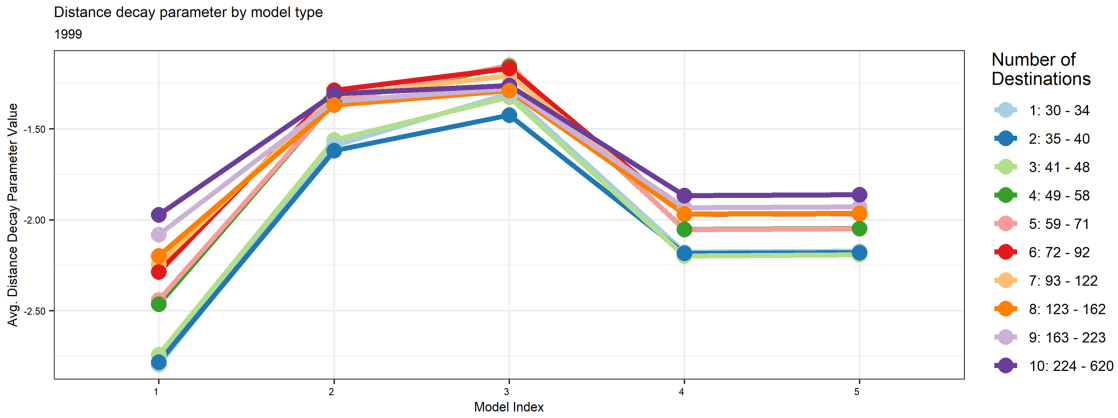


Figure 2.7: Distance decay parameters by model type and destination count groups

coefficients change for counties with many destinations versus fewer destinations. Model 1 and model 2 will produce coefficients with more negative values as these models feature more instances of zero flows. In general, the more observations of zero in this type of model, the more negative the distance decay coefficient. Model 3 will produce coefficients greater than model 1 and model 2 because zeros are removed. Model 4 and model 5 will produce coefficients less than the coefficients produced by model 3 reflecting the inclusion of large county-to-county flows and small county-to-county flows.

Figure 2.7 features the results of the five different counterfactual modelling scenarios and the average coefficient of distance by the grouped destination count. The true distance decay coefficient for each county i is believed to be something similar to the distance decay coefficients produced by models 4 or 5. As seen in figure 2.7, the results of model 1, only including the observed values and known zeros, shifts the coefficients to more negative values. Implementing model 1's configuration would produce coefficients incorrectly indicating that interaction is decreasing more quickly with distance. Models 2 and 3 shift the estimates upwards due to removing the unobserved cells. Implementing models 2 and 3 would produce coefficients incorrectly indicating that interaction is decreasing less quickly with distance. Models 4 and 5 show the effect of assigning a specific value to each of the county-to-county

pairs that can be estimated. Of note, models 4 and 5 produced coefficients that are statistically insignificant from each other but significantly different from models 1, 2, and 3 as evaluated through a paired t-test.

When comparing model 1 with models 4 and 5, as the number of destinations increases, the effect of including the estimated returns on the distance decay parameter decreases. In other words, there are many observations for larger counties and including additional data does not change the coefficients much. It does however change the coefficients for counties with fewer destinations. A goal of this dissertation is to tell the dynamics of migration to and from larger counties and smaller counties. This is only made possible by estimating the suppressed origins and destinations of the aggregate remainder flows (25-percent of all migration data in any given year in the study period). This chapter will now turn to a section on the initial steps to estimate the aggregate remainder IRS county-to-county migration data.

2.5 Identifying and estimating missing flows

Estimating the origins and destinations of the aggregate remainder flows is effectively dealing with missing data due to censoring that was done to prevent disclosure of private tax filling information. In the process of imputing the missing flow data, I identified three types of missing county-to-county flow data:

- Missing flow type 1: Unbalanced incoming and outgoing flows. Flows in the outgoing flow file not recorded in the incoming flow file (or vice-versa).

- Missing flow type 2: Unbalanced within-state flows. Intrastate flows are not completely reported but could be based on an examination of within-state flows.

- Missing flow type 3: Censored flows. A migration flow between any two counties is censored if it did not meet a certain disclosure threshold.

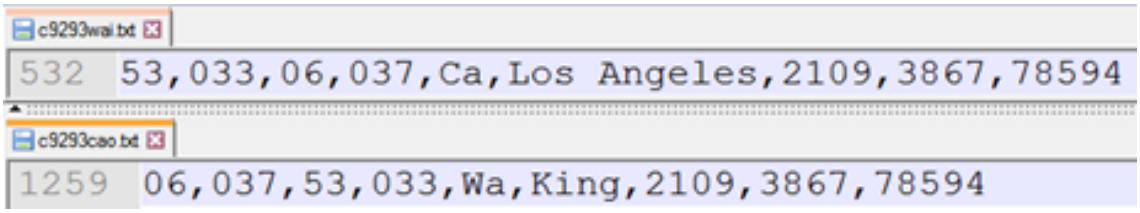


Figure 2.8: Examples of incoming and outgoing flows as presented in the 1992 raw data

2.5.1 Missing Flow Type 1: Unbalanced incoming and outgoing flows

Tabular data analysis is used to shape the reported county-to-county migration flows enabling identification and correction of missing flow type 1. Of the approximately 5.6 million records pertaining to county-to-county migration over the 26-year period, approximately 2.0 million represent county-to-county inflow and approximately 2.0 million represent county-to-county outflow (the remaining 0.6 million pertain to county totals and foreign migration). Given this scheme, the migration data are reported twice: migrants entering a county and migrants leaving a county. Figure 2.8 on page 41 showcases this information for migration from Los Angeles County, California to King County, Washington in 1992 as seen in the raw migration files.

In the incoming file, *c9293wai.txt*, Los Angeles County is listed as the county of origin and 2,109 households migrated to King County. In the outgoing file, *c9293wao.txt*, the same 2,109 households are listed as migrating to King County. In a perfect world, all outgoing migration, if arranged to match the incoming migration data format, would be present in the incoming migration file. Likewise, arranging all incoming flows should match all outgoing flows. This is not the case. For each year of data, while most flows are listed in both the outgoing and incoming file, a small proportion of flows are only listed in the incoming file and a small proportion of flows are only listed in the outgoing file. The technique for identifying and correcting missing flow type is a two-step process. The first step is to arrange all flows to be incoming flows and then remove duplicate flows. And then arrange all flows to be outgoing and then remove duplicates. In total, an additional 1,176 records were added to

Table 2.3: Intrastate migration, Connecticut, 2002

County FIPS	09001	09003	09005	09007	09009	09011	09013	09015	Reported	Remainder	Total
09001		370	444	83	2,065	63	44		3,069	25	3,094
09003	500		726	1,210	1,677	426	1,691		6,230	227	6,457
09005	778	710		47	1,024	27	23		2,609	17	2,626
09007	173	1,408			1,196	368	140	52	3,337	50	3,387
09009	3,926	1,328	795	857		230	106	50	7,292		7,292
09011	115	467		410	256		171	512	1,931	31	1,962
09013	63	2,049		123	110	152		366	2,863	46	2,909
09015	37	204		49	67	599	335		1,291	15	1,306
Reported	5,592	6,536	1,965	2,779	6,395	1,865	2,510	980	28,622		
Remainder			142					269		411	
Total	5,592	6,536	2,107	2,779	6,395	1,865	2,510	1,249			29,033

the incoming flow record set and an additional 1,408 records were added to the outgoing record set. Table B.1 on page 220 in B features the counts of records that were added to each year's data. After this bookkeeping exercise, the total number of incoming records in the 26-year period is 1,981,357. This number is identical to the number of outgoing records. Ensuring that the incoming and outgoing records match sets the stage to identify and correct the second type of missing flow data and then estimate the third type of missing flow data.

2.5.2 Missing Flow Type 2: Unbalanced within-state flows due to inconsistent tabulation

In any state, for all counties, the known intrastate state outgoing flows should sum to the known intrastate incoming totals. For the intrastate migration data, it is only upon estimating the missing within-state flows that the missing flow type is determined: within-state flows that are missing because of inconsistent tabulation or missing within-state flows due to suppression. For some years of data for some states, the reported flows do not sum to the known incoming and outgoing totals. To solve and correct missing within-state flows of this pattern, a linear optimization program is executed to adjust the cell values to ensure that the incoming and outgoing totals sum to equal values. Table 2.3 on page 42 features an example of missing flow data type 2 for Connecticut in 2002.

Table 2.3 depicts the flow of intrastate migration for Connecticut in 2002 arranged in a

matrix. The columns represent the counties of origin and the rows represent the destination counties. The eight counties (New England Townships, technically, but will be referred to as counties) in Connecticut are identified by each county's FIPS code. These eight counties combine to produce 56 possible origin and destination pairs, excluding same origins and destinations (the diagonal in the matrix, shaded grey). Of these possible 56 pairs, 49 have a recorded flow. These 49 county-to-county pairs feature a total of 28,622 reported households moving between counties with an additional 411 households in the aggregate remainder category. Combined, 29,033 returns moved between counties within the state of Connecticut in 2002. The task is to distribute the remaining 411 flows such that the aggregate remainder column and row totals are not exceeded. Certainly, with a bit of math, one could manually distribute the remaining 411 flows[4]. Distributing the 2002 Connecticut within-state returns can also be accomplished using a linear combination of values for each row and each column. In this case, the linear combination of cells is deterministic and there is only one solution for completing the matrix of within-state migration for Connecticut in 2002. The 26-year study period produces 1,300 state-year combinations (this excludes Washington DC as the state of Washington DC is coterminous with the county of Washington DC). Of these 1,300 state-year combinations, approximately 10-percent featured unbalanced intrastate flows.

2.5.3 Missing Flow Type 3: Censored flows

The task of estimating the within-state flows due to censoring becomes more difficult once the number of counties increases or the number of non-reported county pairs increases. For example, there are 254 counties in Texas which combine to produce approximately 64K county-to-county pairs. When considering the number of counties in the US, approximately 3,140 in any year post 1990, there are 9.8M county-to-county pairs in total. Whether the censored flows are intrastate migration or interstate migration, the technique for estimating the origins and destinations of the censored flows is the same: using a combination of spa-

[4]The 411 aggregate remainder flows are trivial and represent a small fraction of within-state moves (1.4-percent of total Connecticut within state moves). It is a convenient example, nonetheless.

Table 2.4: Intrastate migration, Nevada, 2002

County FIPS	32001	32003	32005	32007	32009	32011	32013	32015	32017	32019	32021	32023	32027	32029	32031	32033	32510	Reported	Remainder	Total
32001		23	11	15						62	15	17			105		15	263	21	284
32003	29		73	105			13	12	26	32		397	12		629	44	94	1,466	19	1,485
32005		46								63					181		273	563	31	594
32007		68				21	34	34		11		24			52	17	12	273	13	286
32009																			18	18
32011																			27	27
32013		17		27				17		11			13		44			129	24	153
32015				14			14											28	25	53
32017		50																50	13	63
32019	79	54	115	17								21		15	503		362	1,166	30	1,196
32021	11									12					37			60	25	85
32023		657		12						15					32	11	12	739	39	778
32027															15			15	38	53
32029															256		16	272	24	296
32031	151	588	232	126			106	22		406	24	42	27	214		25	445	2,408	12	2,420
32033		50		13								14			12			89	15	104
32510	20	86	401	14						273				16	329	13		1,152	21	1,173
Reported	290	1,639	832	343		21	167	85	26	885	39	515	52	245	2,195	110	1,229	8,673		
Remainder	44	36	29	21	14	24	38	30	11	15	30	33	18	11	16	12	13		395	
Total	334	1,675	861	364	14	45	205	115	37	900	69	548	70	256	2,211	122	1,242			9,068

tial interaction modeling to provide upper bounds and linear optimization programming to identify a solution with the supplied constraints.

Table 2.4 on page 44 depicts the flow of intrastate household returns for Nevada 2002 arranged in a matrix, similar to the way the data are presented in In Nevada in 2002, there were 395 returns in the aggregate remainder category. Like the county-to-county household flows for Connecticut, estimating the cells in the matrix can be expressed as a linear combination of values. However, solutions for the censored intrastate flows in Nevada in 2002 exist with upper bounds of 6, 7, 8, and 9[5]. Because multiple valid solutions exist using only linear combinations of values, it is necessary to include additional constraints. It is assumed that multiple solutions using only a linear combination of values exist for all censored interstate migration.

2.6 Combining spatial interaction modelling and linear optimization programming

Interstate and intrastate migration in the US is assumed to be a closed system in each year. Given this assumption and the structure of the data, I can state three axioms:

[5]Please see Table C.1 on page 222 and Table C.2 on page 223 for these solutions.

1. For each year, there are county-to-county flows with known origins and known destinations.

2. For each year, for each county, there are counties with censored incoming flows and censored outgoing flows.

3. For each year, for each county, based on the difference between the known county-to-county pairs and all possible county-to-county pairs, there is a set of possible origins and possible destinations for the censored flows.

Remediating missing flow type 1 is a bookkeeping exercise. Finding and remediating missing flow type 2 is possible by expressing the constraints in items one through three in the list above as a series of linear inequalities. While it is possible to find and solve missing flow type 3 by expressing the constraints in items one through three, the solutions generated by the linear optimization program are not guaranteed to be unique. However, it is possible to refine the constraints in a system of linear inequalities by incorporating the distance between every county-to-county pair and estimating the upper bounds of a county-to-county flow using a spatial interaction model. The algorithm involves the following steps: fitting separate spatial interaction models for the within-state and interstate flows to generate a ranking of the county-to-county flows, fitting a regression model to predict the counts of returns of a particular size, and incorporating the ranked data into the linear optimization program as an upper bound for each cell. The linear optimization program is executed to ensure that each cell in a migration matrix is filled according to a set of constraints. Estimating the interstate remainder flows and the within-state remainder flows begins with similar processes. The within-state version of this algorithm identifies missing flow types 2 and 3.

Figure 2.9 features a visual representation of the process flow of the algorithm I have developed. The algorithm is applied to each year's worth of data. There are eight stages for the interstate flows and between eight to twelve stages, for the within-state flows, depending

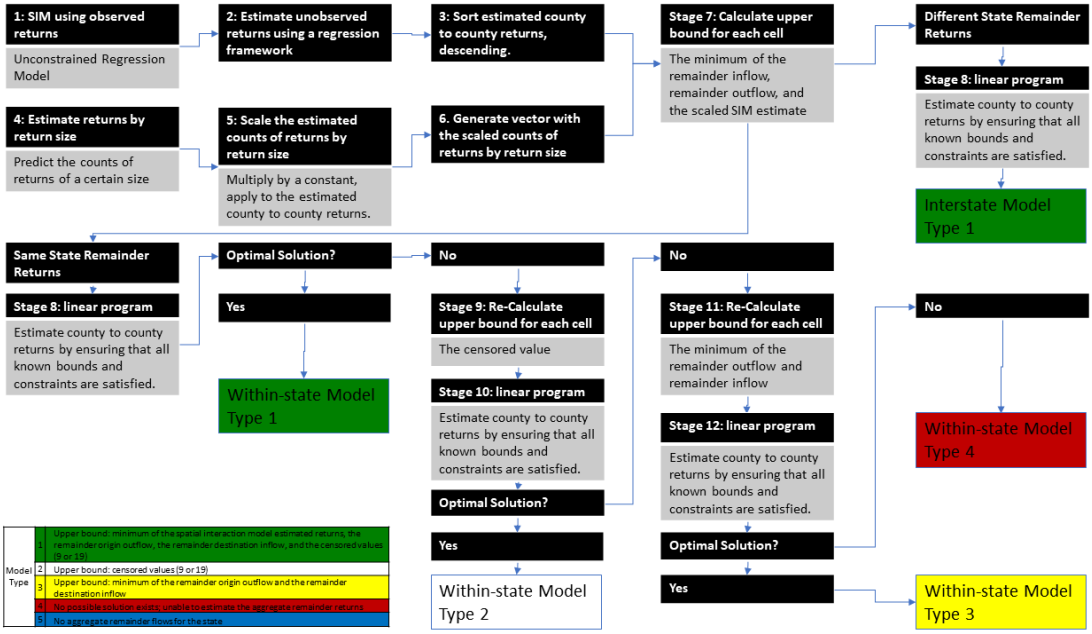


Figure 2.9: Suppressed county-to-county record estimation algorithm flow diagram

on the state and the year. I will describe each stage and conclude this section with diagnostics of the flow estimation algorithm. Where prudent, I will include diagnostics of the individual stages.

2.6.1 Stages 1, 2, and 3: Estimate censored county-to-county returns using the reported county-to-county returns

The first stage in the estimation algorithm is to fit a spatial interaction model (SIM) using the reported returns. This is done separately, for each year, for the within-state and the interstate returns. There is one within-state SIM and one interstate SIM for each year. While chapter three focuses on the techniques and the development of spatial interaction modeling, a brief discussion of spatial interaction modeling is necessary to frame this stage of the algorithm. Spatial interaction models are built on the assumption that the interaction (however broadly defined) between any two areas is a function of each areas' mass (however broadly defined)

and the spatial separation between any two areas (usually defined as the linear distance). This is related to the gravity modeling framework wherein the force between any two bodies is a function of the mass of the two bodies and the distance between the two bodies. Within a migration modeling framework, spatial interaction takes the form of migrants moving from an origin to a destination and each areas' respective mass taking the form of the origin population size and the destination population size. The interaction between the origin and the destination decreases as the spatial separation increases, although the decrease is not always linear. Mathematically, this relationship can be expressed as follows:

$$T_{ij} = K\frac{M_i M_j}{d_{ij}^{\beta}} \tag{2.2}$$

This notational form is taken from Wilson (1971). This is the "classic" Newtonian gravity model and one of its earliest formulations can be traced back to Zipf (1946). In this model, T_{ij} is the flow between areas i and j. Area i is usually considered the origin and area j is usually considered the destination. K is a constant of proportionality to ensure the estimated flows sum to the known observed total, t_{ij}. M_i is the mass term associated with area i and M_j is the mass term associated with area j. Finally, d_{ij} is the distance between i and j and the β value is a term that moderates the degree to which interaction between zones i and j decreases with distance; β is usually a parameter to be estimated. When β, an exponential term, has values close to or equal to one, it means that distance is moderating the relationship between zones i and j linearly. When much less than one, the relationship between zones i and j is decreasing more slowly and when much greater than one, the relationship between zones i and j is decreasing more quickly. In a migration system with a β of 1, two counties that are ten miles apart will have an interaction value equal to 1/10 of the product of county i's mass term and county j's mass term. In a migration system with a β value of 2, two counties that are ten miles apart will have an interaction value equal to 1/100 of the product of county i's mass term and county j's mass term.

Additional parameters tuning the effect of the origin and destination masses can be

included as follows:

$$T_{ij} = KM_i^{\alpha} M_j^{\gamma} d_{ij}^{\beta} \tag{2.3}$$

The exponents to origin i's mass, α, and destination j's mass, γ, increase or decrease the effect of the origin and destination masses. One way of interpreting this is that a relatively larger positive γ value on a destination mass would induce a destination mass to attract like a much larger destination mass. Similarly, values of γ less than one on a relatively smaller destination would make the destination mass behave like an even smaller destination.

The gravity model can be rewritten in multiplicative form:

$$T_{ij} = KM_i^{\alpha} M_j^{\gamma} d_{ij}^{\beta} \tag{2.4}$$

By taking the logarithm of each side and factoring out constants, the following form is obtained:

$$log(T_{ij}) = log(K) + \alpha log(M_i + \gamma log(M_j) - \beta log(d_{ij}) \tag{2.5}$$

This is the log-normal specification of the gravity model and it is possible to calibrate this model using an ordinary least squares regression (Flowerdew 1982). Another way to conceptualize migration within a gravity modelling framework is to note that migration is a non-negative value occurring within a set duration of time.A number of individuals moving between areas i and j will have a Poisson distribution with mean $\lambda_i j$ implying that the probability that n people will move between areas i and j is:

$$Pr\left(T_{ij} = N\right) = \frac{e^{-\lambda_{ij}} \lambda_{ij}^{n}}{n!} \tag{2.6}$$

With the assumption that the λ_{ij} parameter is logarithmically linked to a linear combination of the logged independent variables, the following equation holds:

$$\lambda_{ij} = exp(K + \alpha log(M_i) + \gamma log(M_j) - \beta log(d_{ij})) \tag{2.7}$$

The difference between T_{ij} and λ_{ij} is the realization of the Poisson process. For each year, all intrastate migration was pooled and all interstate migration was pooled. I fit a generalized

linear model, using a Poisson regression for the within-state flows and the interstate flows. Two models were fit for each year for a total of 52 models. For the within-state models, I included a Boolean variable indicating if any two counties are adjacent and in the interstate model I included a Boolean variable indicating if any two counties are adjacent and a Boolean variable indicating if any two states are adjacent. Including the adjacency variables helps control for differences in county and state sizes and shapes. The model specifications are as follows:

$$Within\ state : T_{ij} = M_i + M_j + C_{ij} - D_{ij} \tag{2.8}$$

$$Interstate : T_{ij} = M_i + M_j + C_{ij} + S_{ij} - D_{ij} \tag{2.9}$$

In the above specification, T_{ij} is the number of households that moved from origin county i to destination county j. The M_i and M_j terms are the masses of origin i and destination j: the number of non-moving households. The C_ij term is a Boolean variable indicating if counties i and j are adjacent. In the interstate model, the additional S_ij term is a Boolean variable indicating if county i is in a state that is adjacent to county j. Finally, the D_ij term is the distance in miles between the origin county's centroid and the destination county's centroid. The mass and destination terms are logarithmically transformed on account of the Poisson model specification (Flowerdew 1982).

With the specification of the within-state and interstate models, I can estimate values for the unreported county-to-county pairs. I can also perform a series of model diagnostics to gauge the performance of the models. For each year for both the within-state and interstate models, I randomly split the reported records into two groups: 80-percent of the records went into the training set and 20-percent of the records went into the testing set. Splitting by this ratio ensures that the model is trained on a representative sample (80-percent) and there is enough data not seen by the model (the remaining 20-percent) to gauge model performance (Gholamy et al. 2018). The Poisson model as specified in stage 1 was fit on the 80-percent of

the records in the training set and its performance was evaluated on the held-out 20-percent. All models were significant in each year for both the within-state and interstate models. Table D.1 on page 225 features the R^2 statistics measuring the goodness-of-fit between the reported held-out 20-percent and the fitted values generated from the test data. There is little variation in the R^2) values for the within-state models with an average of 74-percent and range of 72-percent to 77-percent. There is more variation in the interstate models and poorer fit in general with an average of 50-percent and a range of 38-percent to 55-percent. Table D.2 on page 226 and Table D.3 on page 227 feature the coefficients for each within-state model and each interstate model in each year. As I am more concerned with the predictive power of these models and less concerned with the coefficients of the models, I will only note that the coefficients exhibit moderate change year-over-year. For the within-state models, the coefficients exhibit little variation over time. The origin population coefficient ranges from 0.60 to 0.64 with an average of 0.63 and the destination population coefficient ranges from 0.59 to 0.63 with an average of 0.61. The origin and destination population size have equal effects. The distance decay parameter ranges from -0.48 to -0.43 and features an average of -0.46 and the county adjacency variable ranges from 1.50 to 1.67 with an average of 1.59. These two variables are correlated with each other and the inclusion of the county adjacency variable increases the value of the distance decay parameter.

The models estimating the interstate returns exhibit similar patterns over time. The origin population coefficient ranges from 0.63 to 0.70 and feature an average of 0.67 over the 26-year period. The destination population coefficient features an average of 0.67 as well and ranges from 0.62 to 0.71. The distance decay variable ranges from -0.24 to -0.21 and averages at -0.23. The on-average larger interstate distance decay variable is due to interstate returns moving longer distances in addition to the inclusion of the county adjacency and state adjacency variables which average 1.93 and 0.53, respectively. Fitting the within-state and interstate models is stage 1 of the algorithm.

The values for the within-state and interstate remainder returns are predicted given the reported within-state and interstate returns; this is stage 2 of the algorithm. Over the

period in the study, predicting the returns for all remaining county-to-county pairs results in approximately nine times as many returns as there are remaining returns in any given year. And that is to be expected - not all county-to-county pairs are going to feature some form of movement. These models are predicting some amount of household county-to-county movement for all county-to-county pairs. Rather than use these predicted values as the actual flow, I will use these predicted values to generate a ranking of county-to-county pairs. In this context, given the available information, I now have a way of determining which counties should feature a greater share of the censored aggregate remainder returns than others. This ranking scheme is based on a spatial interaction model which incorporates origin and destination population sizes, county distance and county geometry configurations. This technique gives a theoretical and technical definition as to the ranking of the flows between the county-to-county pairs. In practice, this ranking suggests that a flow from San Diego County, California to Colusa County, California will be larger than a flow from San Diego County to Alpine County, California[6]. Once the ranks have been generated, the data are sorted in descending order; this is stage 3 in the algorithm.

2.6.2 Stages 4, 5, and 6: Estimate the counts of returns by return size

Stages 1 through 3 generated a ranking order for each county-to-county pair. In years 1990 through 2012, the maximum unobserved flow between counties is nine and in years 2013 through 2015, the maximum unobserved flow is 19. Censoring is done to protect the privacy and confidentiality of taxpayers and prevent the disclosure of private information. The models specified in stage 3 suggest which county-to-county pairs are going to have a value of nine or eight, but not *how many* county-to-county pairs will feature a flow of nine or eight. I can estimate those counts using the reported returns by assuming that the system-wide total occurrence of a number of returns moving between counties is a function of the size of the

[6]Colusa County, CA is 522 miles northwest of San Diego County and Alpine County, CA is 420 miles north-by-northwest of San Diego County. On average, Colusa County, CA features 19 times as many households as Alpine County, CA.

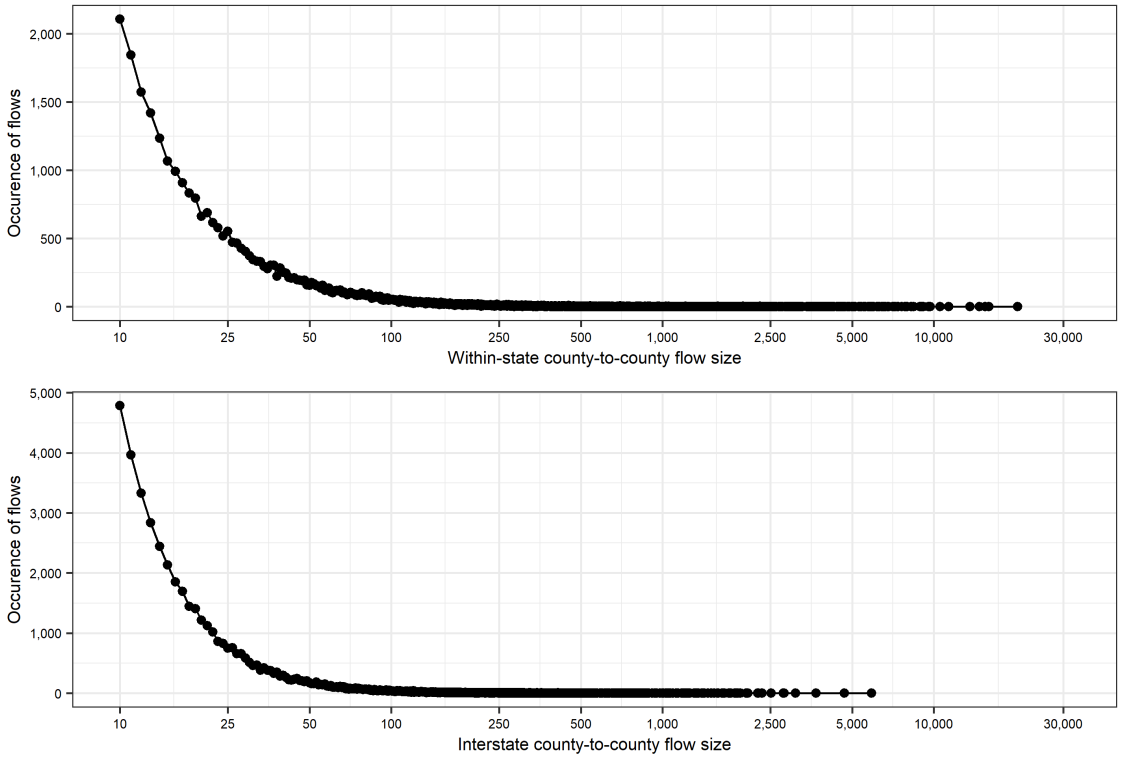


Figure 2.10: Occurrence of within-state and interstate flows by flow size, 1999

flow: the size of the flow between counties predicts the frequency of the flow system-wide. This relationship for both within-state and interstate moves is illustrated in Figure 2.10 on page 52. The two graphics in Figure 2.10 display the relationship between the logarithm of the size of a flow between any two counties and the frequency of that logarithm's occurrence in 1999. For both the within-state and the interstate flows, larger flows are less frequent compared to smaller flows. Using this information, I built a predictive model estimating the number of flows of a certain value.

In this case, the regression model takes the form of:

$$Y \sim f(log(N)) \tag{2.10}$$

This model states that the number of occurrences Y of a flow of size $log(N)$ is a function of the size of N. In this system, smaller flows are more frequent than larger flows. This is

a very simple model, but there is enough data to generate a predictive model. Similar to predicting the flows between counties where I used an 80-20 training-test split, I can do the same for this phase of modeling. Fitting a Poisson model produces an exceptional fit, greater than 99-percent in any given year for within-state returns and greater than 98-percent for the interstate returns. This model enables the calculation of the number of flows of sizes 1 through the maximum censored upper bound (9 for years 1990 through 2012 and 19 for years 2013 through 2015).

2.6.3 Stage 7: Calculate the upper bound of each censored county-to-county pair

Stage 7 involves calculating the upper bound for each censored county-to-county pair by calculating the minimum of three data points for each county-to-county pair. This inequality is expressed as follows:

$$U_{ij} \leq \min\{R_i, R_j, E_{ij}\} \tag{2.11}$$

U_{ij} is the maximum upper bound of a flow between origin county i and destination county j. The value of U for each county i, j pair must be less than or equal to the minimum of the remainder outflow R_i for origin county i, the remainder inflow R_j for destination county j, and the E_{ij} term. The E_{ij} term is the result of combining the ranking of the county-to-county pairs, generated from stages 1 through 3, and the estimated total counts of returns by return size, generated from stages 4 through 6.

2.6.4 Stage 8: Run the linear optimization program

Linear optimization programming is the process of solving a set of linear inequalities, or constraints, so that a minimum (or maximum) value is met (Thapa and Dantzig 1997). Given the data and what is known about the data, I can specify the following constraints:

1. The maximum upper bound of a flow between any two county-to-county pairs without a reported flow is the minimum of the set of outgoing origin county remainder flows,

incoming destination county remainder flows, and the estimated flow. This value is determined in stage 7.

2. For each county, the sum of incoming within-state flows must sum to the reported total number of incoming within-state total flows.

3. For each county, the sum of outgoing within-state flows must sum to the reported total number of outgoing within-state total flows.

These constraints exist for each cell in each state's county-to-county migration matrix and the entirety of the interstate county-to-county migration matrix. Mathematically, the notation is as follows:

minimize:

$$c_1X_1 + c_2X_2 + \cdots c_nX_n = T_{ij}$$

subject to:

$$c_{11}X_1 + c_{12}X_2 + \cdots c_{1n}X_n = b_1$$
$$c_{21}X_1 + c_{22}X_2 + \cdots c_{2n}X_n = b_2$$
$$c_{m1}X_1 + c_{m2}X_2 + \cdots c_{mn}X_n = b_m$$
$$c_n \leq \min\{R_i, R_j, E_{ij}\}$$

In the above maximization problem, T_{ij} is the total number of households migrating within a given state or the number of households migrating across state lines. C_nX_n is the value of the flow from county i to county i $C_1, \cdots, C_n$ is the set of coefficients determined by the linear optimization program to satisfy the above inequality. The values $b_1, \cdots, b_m$ are the total numbers of incoming and outgoing households for a particular county. The maximum value of c_n is determined in stage 7 and is the minimum of the origin county's remainder outflow, the destination county's remainder inflow, and the values estimated by the spatial interaction model.

Once these constraints are entered into the linear optimization program, the program evaluates solutions for each cell until a solution is found that satisfies all constraints. The cells were estimated using Python 3.8 (Van Rossum and Drake Jr 1995) and Gurobi 9.1.1 (Zonghao Gu et al. 2019). For the interstate flows, there is only one outcome: all aggregate remainder interstate flows are a function of the above inequality. For the within-state flows, there are several possible outcomes because there are different state-wide possibilities and the success or failure at subsequent stages represents an outcome dependent upon the types of missing flows for a specific state. If the within-state flows were able to be estimated using the constraints above, then the linear optimization program exits. If that linear optimization program fails, then the constraints are relaxed at most two more times. If all three linear optimization program attempts fail, that state's remainder returns are not used. Any state that does not have aggregate remainder returns in a particular year is marked accordingly.

2.6.5 Stages 9 and 10: Recalculate the upper bound and run the linear optimization program for within-state flows

If a state was not able to be estimated in stage 8, then the constraints are relaxed and the upper bounds take on the following form:

$$c_n \leq \min\{R_i, R_j, C_y\} \tag{2.12}$$

Where C_y is the maximum of the censored value, 9 in years 1990 through 2012 and 19 in years 2013 through 2015. The linear optimization program is executed and if successful, the program exits and the flows for a particular state are recorded.

2.6.6 Stages 11 and 12: Recalculate the upper bound and run the linear optimization program for within-state flows

If the state fails the second round of estimation, the constraints are further relaxed:

$$c_n \leq \min\{R_i, R_j\} \tag{2.13}$$

States that meet this constraint are determined to have missing flow type 2. This is the final stage of constraint relaxation. If this round of constraint relaxation does not work, then that state's returns are not included in that particular year and these additional flows are not included in subsequent analyses. This is relatively rare, only three percent of the 1,300 state-year combinations could not be estimated.

2.6.7 Intrastate return estimation model output

The within-state returns necessitate a different examination tactic when compared to the interstate returns. While a spatial interaction model was used to estimate the upper bound of a county-to-county pair, it was only through the implementation of the estimation algorithm that additional steps in the decision tree were deemed necessary in order to estimate the remainder flows. For each state in each year, there are five possible outcomes for the estimation of the aggregate remainder returns where each subsequent outcome features fewer constraints than the previous. Figure 2.11 features the color-coded outcome of the within-state aggregate remainder return estimation algorithm. Color coding was implemented to showcase trends and relative proportions of linear optimization program output. The list below features the five possible outcomes based on the upper bound for each within-state county-to-county origin-destination pair:

1. The minimum of the spatial interaction model estimated returns, remainder origin outflow, remainder destination inflow, censored values (9 or 19): cells marked in green. If successful, this represents stage 8 of the linear optimization program.

2. The upper bound of the censored values (9 or 19): cells marked in white. If successful, this represents stage 10 of the linear optimization program.

3. The minimum of the remainder origin outflow and the remainder destination inflow: cells marked in yellow. If successful, this represents stage 12 of the linear optimization program and represents estimating missing flow type 2.

4. If the remainder returns were not able to be estimated for a state, then no possible within-state solution exists: cells marked in red.

5. No aggregate remainder flows for the state: cells marked in blue.

Across the 26-year period, there are 1,300 state-year combinations. I was able to estimate the origins and destinations of the aggregate remainder flows using model type 1 for 19-percent of the state-year combinations while 65-percent of the state-year combinations needed the relaxed constraints as found in model type 2. Approximately 10-percent of the state-year combinations were estimated using model type 3 while 3-percent of the state-year combinations could not be estimated and two percent of the state year combinations did not feature aggregate remainder returns. For the state year combinations that were not able to be estimated, a total of approximately 13K returns (less than 1/1000 of a percent of all-years returns) across the 26-year time are not included in subsequent analyses.

2.6.8 Interstate return estimation model output

As the interstate models have only one model exit status, I can diagnose the values estimated by the interstate linear optimization program to understand its performance. The estimation algorithm was designed in such a way as to ensure that only a certain number of county-to-county pairs were assigned values of 1 through 9 for years 1990 through 2012 and values 1 through 19 for years 2013 through 2015. However, not all county-to-county pairs that featured a estimated value were assigned a value. If a county-to-county pair was assigned a value, over 99-percent of the time it was assigned a value of zero. The three graphs in Figure 2.12 on page 60 feature the scaled counts of records - county-to-county pairs - that feature a maximum flow size of 1 through 19, graphic A, the counts of the county-to-county

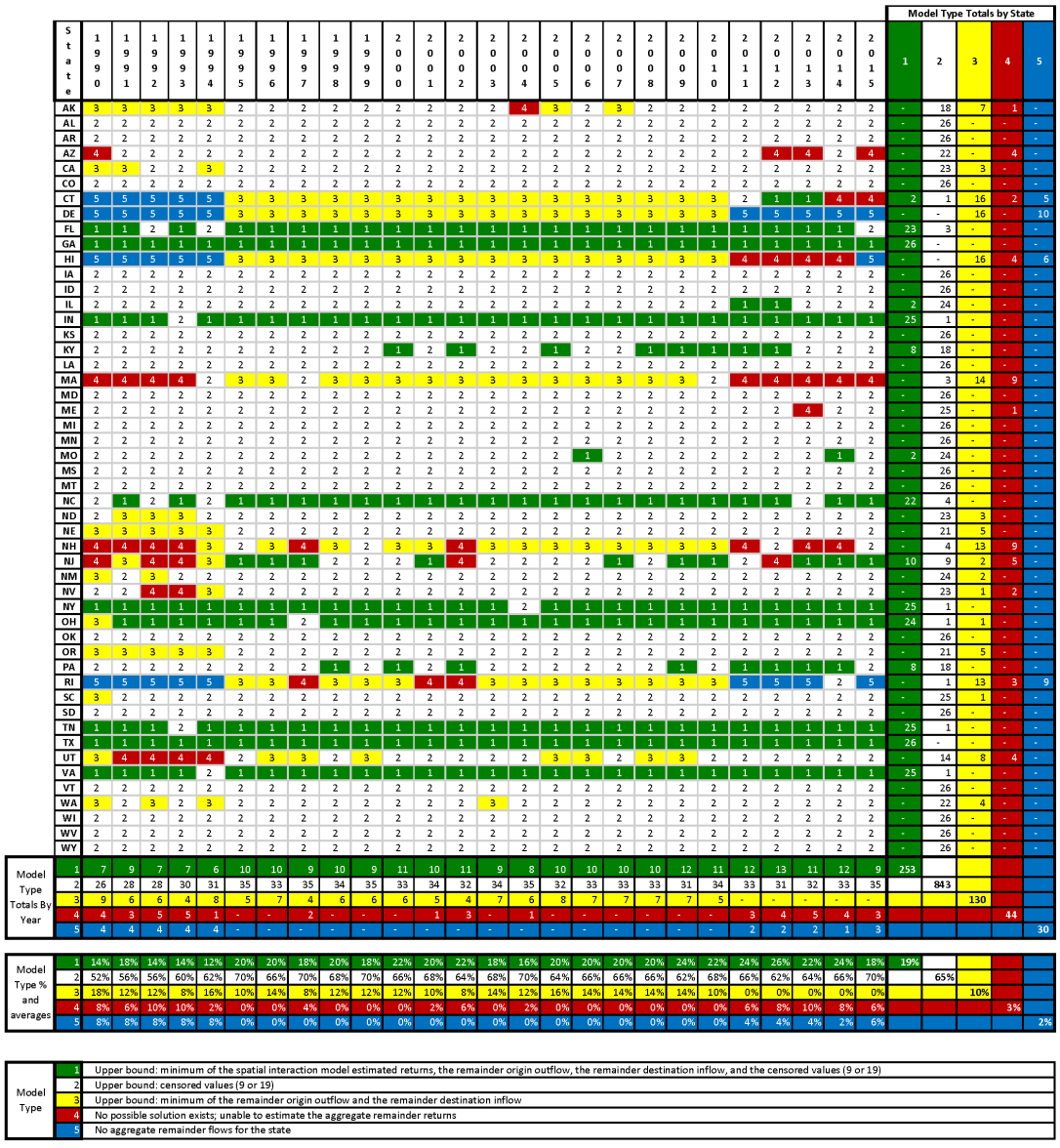


Figure 2.11: Within-state linear optimization program solution status by state and year, 1990-2015

pairs that were assigned a flow of 1 through 19, graphic B, and the proportion of county-to-county pairs that were assigned a flow of 1 through 19, graphic C.

Graphic A in Figure 2.12 shows that the pre-LP-estimation count for all years, most county-to-county pairs was assigned a value of one, followed by two, followed by three, etcetera. Graphic A features the pre-LP-Estimation counts; these are the counts of county-to-county pairs by the maximum flow size. The values on display in graphic A follow from the assumption that the counts of flows by flow size originate from a Poisson distribution. As the IRS's data reporting techniques changed in 2013, the average counts of county-to-county pairs by flow size changed in beginning in 2013. Graphic B of Figure 2.12 showcases the post LP-estimation counts of county-to-county pairs by flow size. For years 1990 through 2012, the counts are consistent and follow the trends on display in graphic A. In years 2013 through 2015, the counts shift due to the change in the IRS's data reporting techniques. Finally, graphic C in Figure 2.12 showcases the proportion of the pre-LP-Estimation counts of county-to-county pairs by flow size that were assigned a value other than zero during the LP-Estimation process. On average, in each year in the 1990 through 2012 period, of the approximately 2.7M county-to-county flows that featured a maximum constraint of one, approximately 251K or 9-percent of the county-to-county pairs received a value of 1. The remaining 91-percent of the 2.7M county-to-county flows received a value of 0. Similarly, in each year of study period, approximately 141K county-to-county pairs featured a pre-LP-estimation maximum constraint of four returns. Of these 141K county-to-county pairs, 36K pairs featured a post-LP-estimation value of 4 returns, representing 25-percent of the total possible. The remaining 105K returns were assigned a value of zero. The proportions of pre-LP-estimation counts are consistent during the 1990 through 2012 period. The proportions change in years 2013, 2014, and 2015, however.

Figure 2.12 illustrates that only a proportion of the county-county pairs that featured a pre-LP-estimation value were assigned that value. This suggests the question that if only a small proportion of the county-to-county pairs featuring a maximum constraint of 7 (or 3 or any other positive, non-zero number) were assigned a value of 7 (or 3 or any other positive,

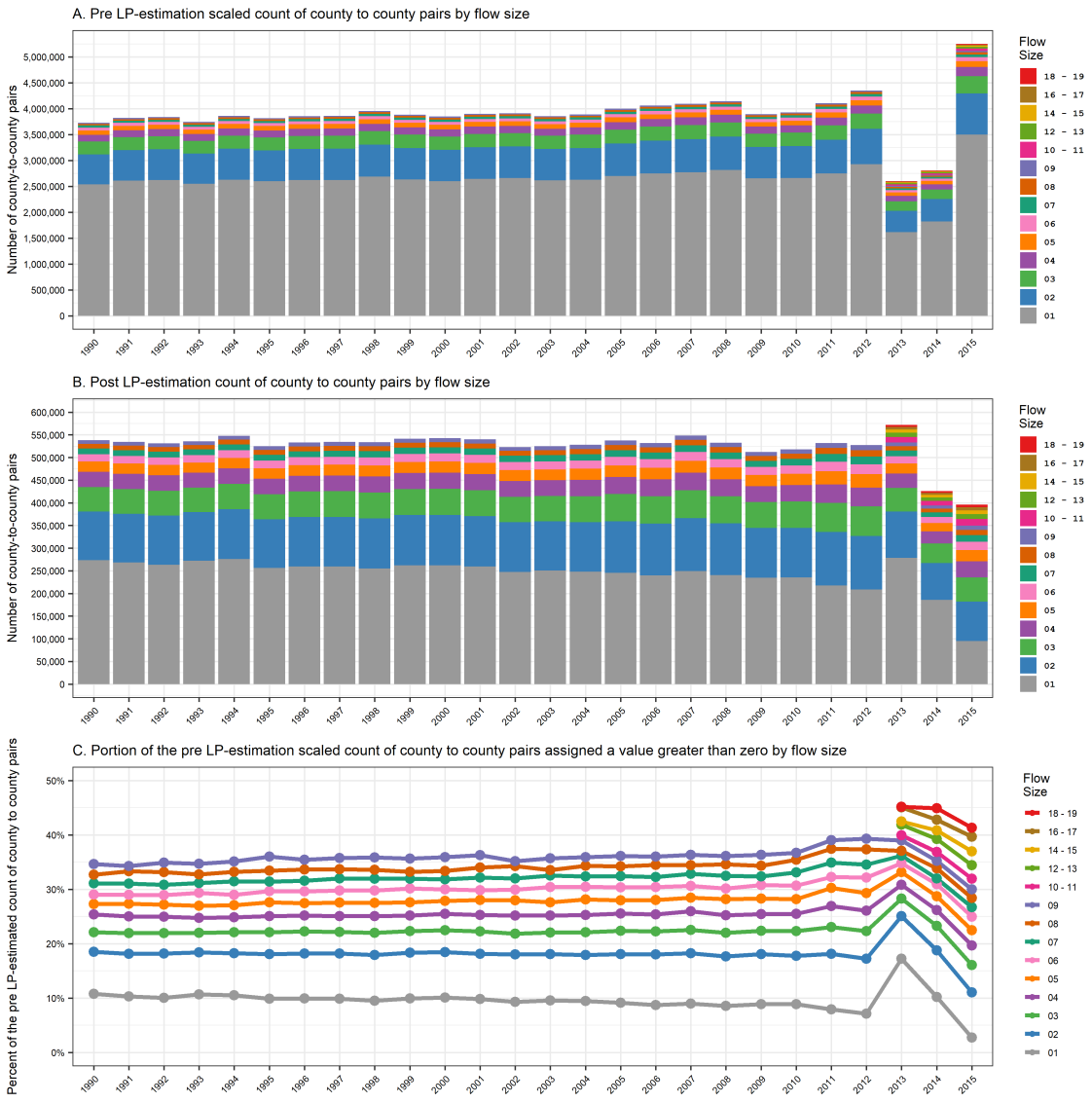


Figure 2.12: Analysis of the estimated values of records by flow size, 1990-2015

non-zero number), why were particular county-to-county pairs assigned a value greater than zero while others were not? This question can be answered by modeling the outcome of the LP-estimation part of the algorithm as a Boolean variable: a county-to-county pair was either assigned a value greater than zero or it was assigned zero.

2.6.9 Modeling county-to-county pair inclusion

To illustrate why a county-to-county pair was assigned a value greater than zero, I fit a logit model with the following reduced form:

$$logit(V_{ij}) = X\beta \tag{2.14}$$

Here, the vector of explanatory variables, X, is calculated from the same data that was used in the linear optimization programming part of the censored flow data estimation algorithm. The dependent variable, V_ij, indicates if county-to-county pair ij was assigned a value of zero (0) or a value greater than zero (1). The explanatory variables generated for each year in the study period are listed in Table 2.5 on page 62. These variables were based on what I input into the linear optimization program; no additional data were incorporated in this phase of the analysis. To test the performance of the model, I used an 80-20 train-test split for each year of data to see how well my model performed. The model was trained on 80-percent of the data and tested on the remaining 20-percent. Table D.5, page 229, features the coefficients from the model for each year. As I am more concerned with the models' explanatory performance rather than the coefficients, I will forego discussing the coefficients though I will note that all models are significant at the 0.05 level and nearly all of coefficients are significant for each year. After predicting the values for the held-out 20-percent, I prepared a confusion matrix to judge model accuracy. A confusion matrix determines the rate at which the model estimates true positives (1,1), false positives (0,1), true negatives (0,0), and false negatives (1,0). Table D.6 on page 230 features a confusion matrix depicting model accuracy for each year. On average, for years 1990 through 2012, the model is approximately 86-percent accurate when predicting true negatives and true

Table 2.5: Explanatory variables used in the county-to-county pair selection model

Description
The upper bound of the county-to-county pair
The number of destination pairs for the destination county
The number of origin pairs for the origin county
The number of instances of the destination county with the particular upper bound
The number of instances of the origin county with the particular upper bound
The destination county's remainder inflow averaged over the number of possible origins
The origin county's remainder outflow averaged over the number of possible destinations
The destination county's total inflow averaged over the number of possible origins
The origin county's total outflow averaged over the number of possible destinations
The ratio of a county's incoming remainder flow to its total incoming flow
The ratio of a county's outgoing remainder flow to its total outgoing flow
The ratio of the destination's incoming remainder returns to the origin's outgoing remainder returns
The ratio of the destination's incoming total returns to the origin's outgoing total returns

positives. In years 2013, 2014, and 2015 the model's accuracy was 76-percent, 84-percent, and 92-percent, respectively.

These logistic regression models showcase why careful selection of the upper bounds of a county-to-county migration pair is necessary. The linear optimization program distributes censored values by fulfilling a set of constraints: meeting each county's total number of incoming and outgoing households and the upper bound of a flow between county-to-county pairs. By implementing a combination of spatial interaction modeling and linear optimization programming, I was able to set the upper bounds of the county-to-county pairs. The diagnostic logistic regression models indicate the appropriateness of the computed upper bound of the flow between each county-to-county pair. The logistic regression model predicting greater rates of false positives or false negatives would indicate that the computed upper bounds of the flows between county-to-county pairs were not useful.

This concludes the description and the evaluation of the algorithm I implemented to estimate the within-state and interstate aggregate remainder returns. The algorithm involved a combination of spatial interaction modeling and linear optimization programming to estimate the origins and destinations of the remainder returns. Having implemented this

algorithm, I can now prepare statistics and tabulations describing the combined reported and estimated county-to-county household migration data in relation to just the reported county-to-county household migration data.

2.7 Enhanced county-to-county migration data: combining reported and estimated county-to-county flows

Combining the reported county-to-county household migration flows and the estimated county-to-county household migration flows produces the enhanced county-to-county household migration dataset. I can showcase four aspects illustrating the impact of including the estimated flows with the reported county-to-county migration data: the number of additional county-to-county pairs with a household flow, the number of counties with at least 30 origins / destination, returns by county adjacency distance, and returns by metropolitan status. The first aspect of the enhanced data I will examine is the number of reported county-to-county records and estimated county-to-county records.

Table 2.6 on page 64 features the reported and estimated counts of the county-to-county records by within-state and interstate state. I was able to estimate an additional 50K-68K within-state records per year during the 1990-2015 period which increases the number of within-state records by nearly three-fold. The interstate returns exhibit a similar pattern across the years, but with different magnitudes. For years 1990 through 2013, I was able to estimate an additional approximately 512K-548K interstate records per year, representing a 10-to-13-fold increase in the number of interstate county-to-county records with a flow. In 2013, 2014, and 2015, the additional interstate records represent a 25-, 25-, and 17-fold increase in the number of interstate county-to-county records with a flow.

As additional county-to-county pairs have now been estimated, I now have more counties that feature at least 30 origins or 30 destinations. The number of unique origins or destinations per county is important for statistical significance within a regression framework. Given the heuristic of 10 observations per variable in a regression model (Troutt 2006), and the variables I will select for the spatial interaction modelling in chapter three, I am interested

Table 2.6: Reported and estimated county-to-county pairs, 1990-2015

Year	Within-State Reported	Within-State Estimated	Within-State Total	Within-State % Increase	Interstate Reported	Interstate Estimated	Interstate Total	Interstate % Increase
1990	30,698	50,271	80,969	264%	43,763	537,883	581,646	1329%
1991	30,922	54,407	85,329	276%	43,568	534,268	577,836	1326%
1992	31,137	52,372	83,509	268%	43,982	531,092	575,074	1308%
1993	31,306	52,676	83,982	268%	43,695	535,807	579,502	1326%
1994	31,815	49,759	81,574	256%	45,372	547,884	593,256	1308%
1995	31,872	61,162	93,034	292%	44,594	525,088	569,682	1277%
1996	32,643	61,351	93,994	288%	45,668	533,049	578,717	1267%
1997	32,802	60,356	93,158	284%	46,189	534,516	580,705	1257%
1998	33,428	61,863	95,291	285%	46,832	533,790	580,622	1240%
1999	33,711	60,450	94,161	279%	47,229	541,300	588,529	1246%
2000	33,791	66,379	100,170	296%	47,837	542,521	590,358	1234%
2001	33,862	60,789	94,651	280%	47,337	539,821	587,158	1240%
2002	33,398	64,927	98,325	294%	45,919	523,250	569,169	1240%
2003	33,338	59,203	92,541	278%	46,042	524,656	570,698	1240%
2004	33,651	59,373	93,024	276%	48,068	527,772	575,840	1198%
2005	34,733	64,706	99,439	286%	50,646	537,294	587,940	1161%
2006	34,039	63,654	97,693	287%	49,672	531,935	581,607	1171%
2007	35,005	61,108	96,113	275%	50,634	548,525	599,159	1183%
2008	34,396	64,428	98,824	287%	48,766	532,184	580,950	1191%
2009	32,977	63,100	96,077	291%	45,402	512,347	557,749	1228%
2010	33,358	62,661	96,019	288%	47,834	517,721	565,555	1182%
2011	35,147	67,404	102,551	292%	54,379	531,882	586,261	1078%
2012	35,431	68,210	103,641	293%	56,028	527,647	583,675	1042%
2013	21,528	59,689	81,217	377%	23,920	571,625	595,545	2490%
2014	18,439	61,979	80,418	436%	17,562	425,743	443,305	2524%
2015	21,906	61,450	83,356	381%	24,983	395,815	420,798	1684%

in working with counties with at least 30 origins or destinations.

Table 2.7 on page 66 features the counts of counties with at least 30 origins or destinations. For years 1990 through 2012, approximately 17-percent of counties (540 per year) featured at least 30 origins or destinations. After including the estimated county-to-county pairs, approximately 95-percent of counties (3K per year) feature at least 30 origins or destinations. For years 2013, 2014, and 2015, the included estimates bring the proportions of counties with at least 30 origins or destinations to 91-percent, 87-percent, and 88-percent, respectively. This large increase in the number of counties with more unique origins and destinations enables a more thorough analysis of the spatialities of internal migration.

2.8 Revised yearly trends in internal migration, 1990-2015

Figure 2.6 on page 36 features the yearly totals of within-state and interstate returns disaggregated by reported and aggregate remainder status. While the total number of returns will not change after including the estimated returns[7], I can now prepare statistics comparing the combined reported and estimated flows with just the reported flows. Figure 2.13 on page 67 features incoming returns by county adjacency distance. County adjacency distance is operationalized as the adjacency distance between counties. Any two counties are considered adjacent if they share a border. For example, King County, Washington is directly adjacent to Pierce County, Washington, to the south and Snohomish County, Washington, to the north. Both Pierce and Snohomish counties have an adjacency distance of 1 from King County. As Pierce and Snohomish counties do not share a border, each county has an adjacency distance of 2 from the other. Adjacency distance is used as opposed to a linear distance measurement (such as miles) to illustrate how households have moved by county[8].

Figure 2.13 on page 67 features four separate graphs combined into one layout. Graphics

[7]Except for the 13K within-state returns that I could not estimate.

[8]F features two maps: Figure F.1 on page 237 and Figure F.2 on page 238 depicting the counties directly adjacent to Clark County, Nevada and DeKalb County, Georgia, respectively. A move from Clark County, Nevada to an adjacent county is on average of 115 miles while a move from DeKalb County, Georgia to an adjacent county is approximately 17 miles on average.

Table 2.7: Number of counties with at least 30 origins or 30 destinations, reported and estimated data, 1990-2015

	Counties with at least 30 origins				Counties with at least 30 destinations			
Year	Reported	Estimated	Total	% Increase	Reported	Estimated	Total	% Increase
1990	516	2,554	3,070	495%	501	2,589	3,090	517%
1991	517	2,559	3,076	495%	499	2,592	3,091	519%
1992	520	2,560	3,080	492%	491	2,593	3,084	528%
1993	525	2,551	3,076	486%	497	2,585	3,082	520%
1994	535	2,548	3,083	476%	512	2,574	3,086	503%
1995	534	2,445	2,979	458%	512	2,467	2,979	482%
1996	523	2,451	2,974	469%	534	2,464	2,998	461%
1997	529	2,449	2,978	463%	541	2,457	2,998	454%
1998	543	2,422	2,965	446%	547	2,462	3,009	450%
1999	542	2,419	2,961	446%	548	2,458	3,006	449%
2000	542	2,426	2,968	448%	554	2,450	3,004	442%
2001	545	2,423	2,968	445%	544	2,448	2,992	450%
2002	547	2,417	2,964	442%	527	2,459	2,986	467%
2003	541	2,413	2,954	446%	530	2,452	2,982	463%
2004	554	2,404	2,958	434%	554	2,426	2,980	438%
2005	584	2,383	2,967	408%	569	2,407	2,976	423%
2006	581	2,385	2,966	410%	567	2,403	2,970	424%
2007	586	2,395	2,981	409%	577	2,400	2,977	416%
2008	562	2,419	2,981	430%	564	2,408	2,972	427%
2009	526	2,453	2,979	466%	539	2,414	2,953	448%
2010	534	2,444	2,978	458%	562	2,410	2,972	429%
2011	588	2,373	2,961	404%	599	2,339	2,938	390%
2012	598	2,374	2,972	397%	615	2,333	2,948	379%
2013	319	2,543	2,862	797%	333	2,547	2,880	765%
2014	256	2,469	2,725	964%	261	2,531	2,792	970%
2015	337	2,421	2,758	718%	343	2,431	2,774	709%

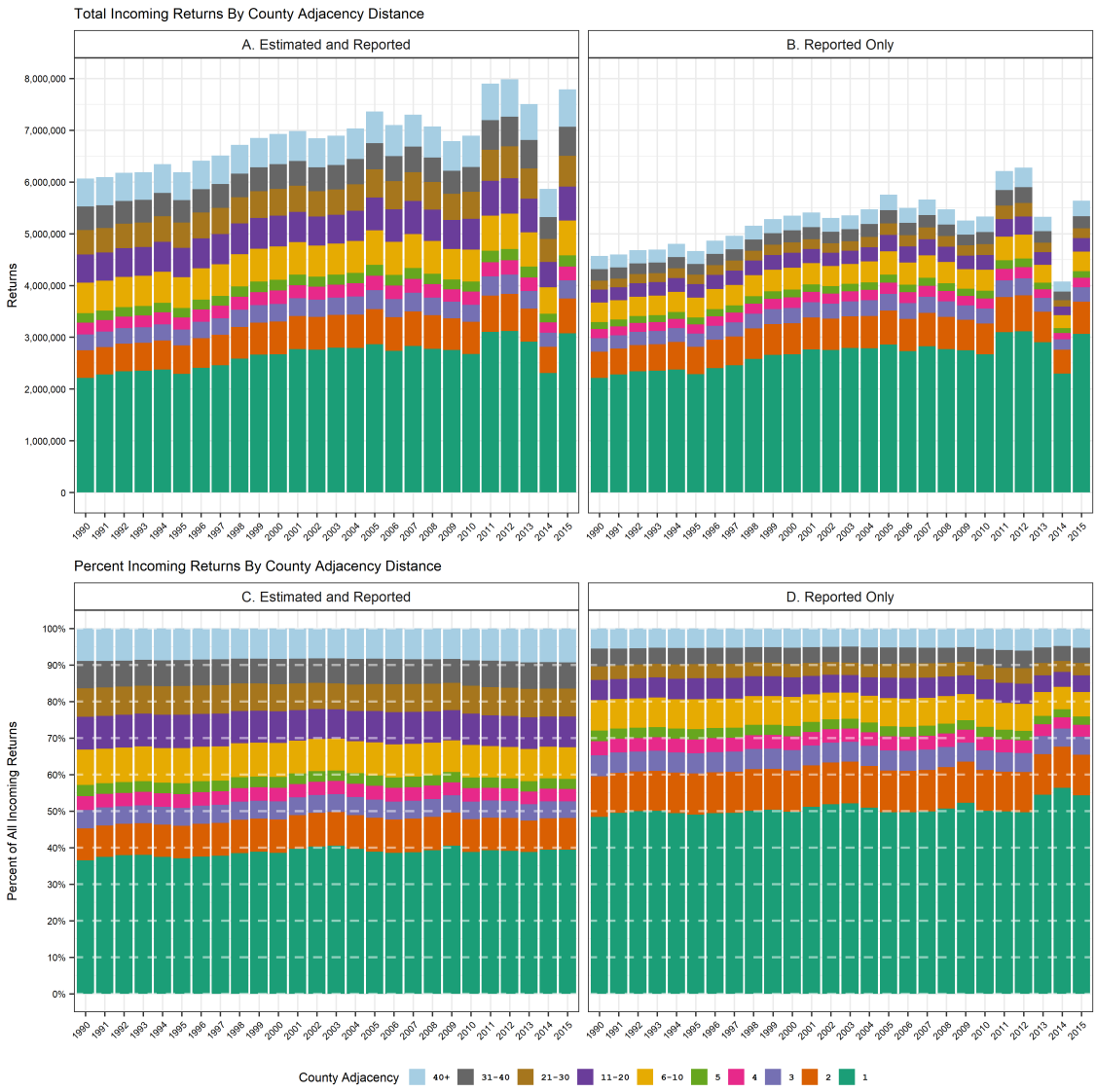


Figure 2.13: Incoming returns by county adjacency distance, 1990-2015

A and C feature the combined reported and estimated county-to-county returns and graphics B and D feature only the reported county-to-county returns. Graphics A and B feature counts by adjacency distance and the graphics C and D express those counts as a percentage of all county-to-county movement.

The inclusion of an additional approximately 1.6 million county-to-county returns per-year is for county-to-county pairs that have an adjacency distance of two or more. Most reported county-to-county movement is from one county directly adjacent to another county and this is featured in the reported data. Using only the reported returns, in any year in the study period, approximately 50-percent of moves are to a directly adjacent county. Including the estimated county-to-county returns lowers the percent of moves to a directly adjacent county to approximately 35-percent in any given year in the study period.

Figure 2.14, page 69 - the second graphic showcasing the effect of including the estimated returns - features four graphics depicting movement between rural and metropolitan areas. Graphics A and B in Figure 2.14 feature the combined reported and estimated returns and graphics C and D feature the reported returns only. As in Figure 2.13, graphics A and C feature counts and graphics B and D feature the counts expressed as a percentage. In addition, Figure 2.14 showcases whether the move is to a directly adjacent county or a county that is not adjacent. The metropolitan data come from the Office of Management and Budget and the US Census Bureau US Census Bureau (2016, 2013, 2010, 2009, 2007, 2006, 2005, 2004, 2003, 1999, 1993, 1990).

Graphic B in Figure 2.14 showcases how, since 1990, a larger proportion of returns are moving within and between metros. In 1990, approximately 65-percent of all moves were within and between metros and in 2015 that proportion had increased to approximately 75-percent. Consequently, the share of moves between non-metropolitan areas, out of a metropolitan area, and to a metropolitan area decreased. Chapter four discusses non-metropolitan-to-non-metropolitan, non-metropolitan-to-metropolitan, metropolitan-to-non-metropolitan, metropolitan-to-metropolitan moves in greater detail. Using only the reported returns suggests that internal migration trends remained steady during 1990 to 2016 period.

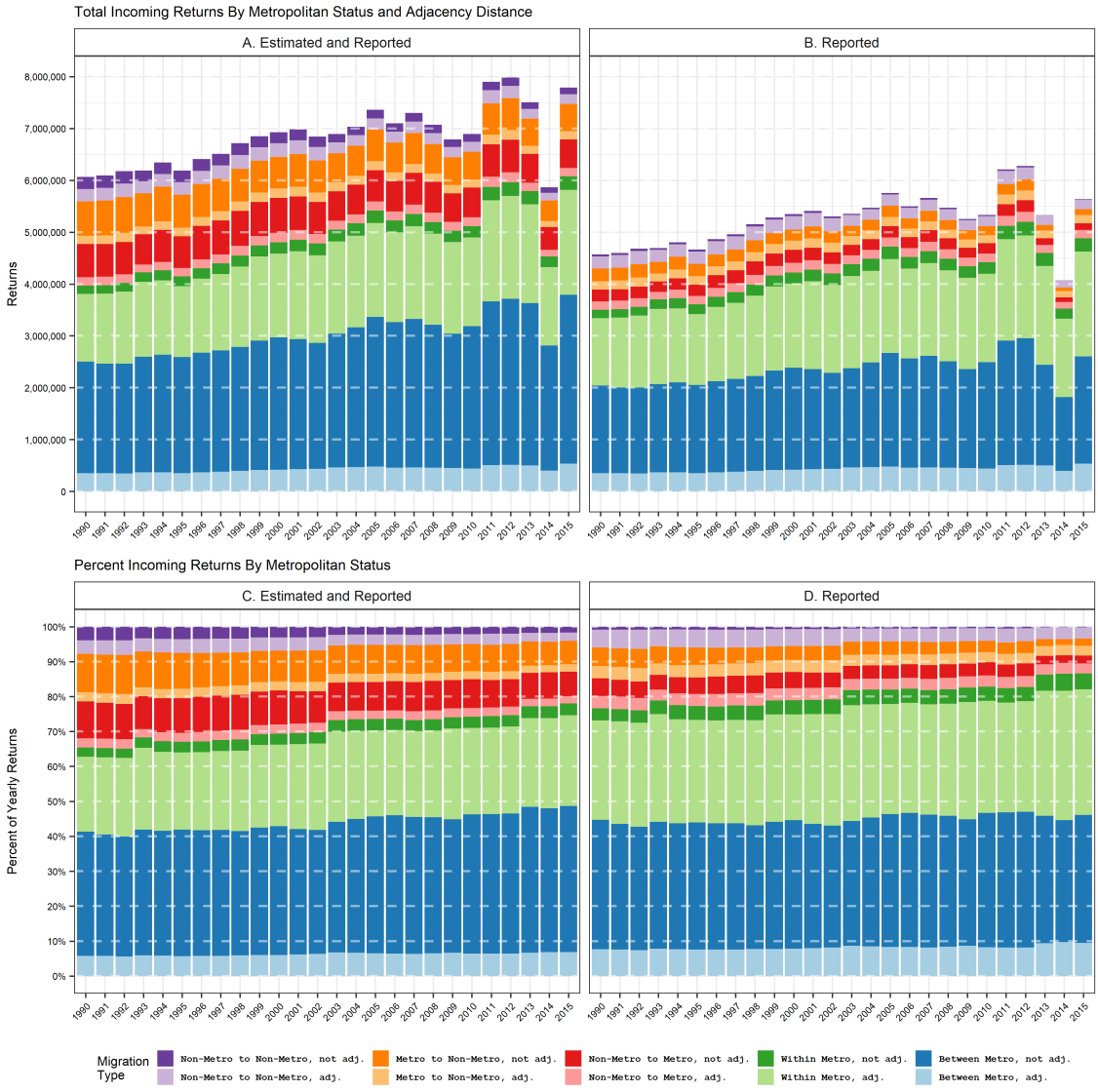


Figure 2.14: Number and incoming migration rate by metropolitan status and adjacency distance, 1990-2015

The inclusion of the estimated returns, however, suggests that internal migration in the US is focusing on metropolitan migration.

2.9 Enhancing IRS county-to-county household migration data

I describe six tasks in this chapter that showcase the IRS county-to-county household migration data and the algorithm I developed to enhance the IRS county-to-county household migration data. First, I described the IRS county-to-county household migration data and second, a database schema suitable for all years of the IRS county-to-county migration data. Third, I described the internal migration trends present in the county-to-county migration data. Fourth, I described the identification and estimation of the remainder returns. Fifth, I described the algorithm I wrote to estimate the remainder returns and finally, I described the results of the implementation of the estimation algorithm by combing the reported and the estimated returns. The first task described in greater detail the data I am working with: The IRS's county-to-county Migration data for the years 1990 through 2015. The migration data do not feature a consistent schema and initial analysis proved difficult. The second task described in this chapter involved harmonizing these data over the 26-year period and placing the data into a schema that enabled the rapid and accurate identification of records for county-to-county migration analysis.

After instantiating a suitable database schema, I then described some of the trends in the migration data via within-state, interstate, and foreign migration categories. I then disaggregated the within-state and interstate migration data to focus on the reported and remainder within-state and interstate migration data. Flows between counties are fully reported when the size of the flow clears a threshold: 10 or more households in years 1990 through 2012 and 20 or more households in years 2013, 2014, and 2015. Returns between counties that do not clear the disclosure threshold are placed in an aggregate remainder within-state category or an aggregate remainder interstate category. This aggregation technique is done to protect taxpayer's privacy and confidentiality. As result of this suppression, in any given year, only 75-percent of the data are fully reported. I demonstrated the utility of estimating

the remainder returns by showing the effect of including additional records on a distance decay coefficient originating from a spatial interaction model. Including additional records removes bias in regression parameter estimates.

The fourth task in this paper described three types of missing county-to-county flows and detailed the technique used to estimate these missing flows. The fifth task featured the description and implementation of an eight-to-twelve step algorithm that estimates censored within-state and interstate remainder flows using a combination of spatial interaction modeling and linear optimization programming. The sixth section featured an examination of the output of the estimated county-to-county flows and the reported county-to-county flows with only the reported flows. This combined reported and estimated flow data is the enhanced IRS county-to-county migration dataset. The implementation of the remainder flow estimation algorithm resulted in an average increase of 68K within-state records and an average increase of 525K interstate records per year.

The inclusion of the estimated returns increases the proportion of households moving longer distances. Using only the reported household migration data, the proportion of households moving to a directly adjacent county is approximately 50-percent. By incorporating the remainder returns the proportion of households moving to adjacent counties is approximately 35-percent in any given year. Using only the reported data and a county's metropolitan status, moves between and within metros are approximately 75-percent of moves in 1990 and gradually increase to approximately 85-percent of moves by 2015. The inclusion of the estimated data and a county's metropolitan status illustrates that moves between and within metros are approximately 65-percent of moves in 1990 and gradually increase to 75-percent of moves by 2015. The inclusion of the estimated data better illustrates the changes in moves to and from metropolitan areas and between non-metropolitan areas and enables a more thorough analysis of county-to-county household migration in the United States.

Chapter 3

THE DESTINATION PREFERENCES OF HOUSEHOLD MOVEMENT: "I'VE MET MORE PEOPLE FROM CHICAGO THAN I HAVE FROM OKANOGAN."

While previous studies have shown migration rates are in a period of decline (Cooke 2013; Molloy et al. 2011) an investigation into the spatiality of migration patterns under this decline remains to be undertaken. This chapter investigates the spatiality of migration patterns during the 1990 through 2015 period by examining four destination preferences of households: population size, centrality, distance, and interstate movement. Using reported and estimated county-to-county household migration data, I fit origin-specific, production-constrained, spatial interaction models for each county for each year during the 1990 through 2015 time-period for a total of approximately 77K spatial interaction models. The spatial interaction models feature the same specification with the following four covariates: the destination mass size, the distance between each origin and destination, the accessibility of the destination, and a variable indicating if the move is within the same state of origin or not. These models produce yearly origin specific regression coefficients for population size, distance, accessibility, and state boundaries that provide insight into the changing dimensions of households preferences over time. As the data driving this chapter feature a sizable temporal and spatial dimension, multiple visualization techniques are used to showcase concisely the spatiality of migration preferences.

This chapter begins with a description of spatial interaction models and the second section features a description of the exact spatial interaction model specification I will use. The third section describes the data I am using for my models and the fourth section showcases the results of the models. The fifth and final section concludes with a discussion of next steps.

This chapter enumerates some of the broad migration trends seen over a 26-year period with a substantial geographic component: county-to-county household migration in the US from 1990 through 2015. Accordingly, there are two goals of this chapter. The first is to illustrate how spatial interaction models can be used with the reported and estimated county-to-county IRS household migration data and the second goal is to describe how migration trends can be ascertained from a small number of variables from two sources of data.

3.1 Spatial interaction models

A spatial interaction model encapsulates interaction between an origin and a destination as a function of the origin's mass, the destination's mass, and the distance between the origin and destination. This relationship between mass and distance is exemplified by migration into King County, Washington from Cook County, Illinois and Okanogan County, Washington. Cook County is approximately 1700 miles from King County and is the central county in the Chicago metropolitan area. Okanogan County is approximately 120 miles northwest of King County. Cook County has approximately 150 times the population of Okanogan County. The size of the flows originating in Cook County and terminating in King County are ranked 15th in 1990, 11th in 2000, and 9th in 2015. Flows originating in Okanogan County and terminating in King County are ranked 79th in 1990, 88th in 2000, and 148th in 2015. Even though the distance from King County to Cook County is 14 times the distance of King County to Okanogan County, in 2015, approximately 19 times as many households moved from Cook to King County when compared to the number of households moving from Okanogan to King[1]. This chapter is focused on the relationships between origins, destinations, and the distances between them. A simple way of expressing this relationship mathematically is:

$$T_{ij} = \frac{X_i X_j}{D_{ij}} \tag{3.1}$$

[1]To date, while I have met many people from Chicago (Cook County, IL), I have only met one person from Okanogan County (and she lived in Chicago for several years!).

Equation 3.1 equates T_{ij}, the total interaction between area i and area j, as a combination of a vector of origin specific characteristics X_i, a vector of destination specific characteristics X_j and the geographic distance D, between i and j. The form of the spatial interaction model in equation 3.1 is often referred to as a Newtonian gravity model. The nomenclature of the model is derived from the notion that an origin i has some degree of emissivity, operationalized by the X_i vector and a destination j has some degree of attractivity, operationalized by the X_j vector. Early formations of models of this type can be seen in the 1940s. Zipf (1946, 1949) investigated the movement of people between cities using the population of the origin and destination as values for the X_i and X_j terms and therefore equating movement between cities as a function of the mass of each city and the distance between the cities. Zipf's model assumed that the effect of distances between cities was linear.

Many scholars pushed the development of the spatial interaction model throughout the 20th century. Stouffer (1960, 1960), citing Ravenstein (1889, 1885), suggested that more moves are of shorter distance, and fewer moves are of longer distance and proposed the theory of intervening opportunities. This concept treated the number of flows emanating from an origin as being inversely proportional to the number of opportunities a given distance from an origin. Stouffer's focus on distance is a reformulation of a gravity model and treated distance as a determining factor. Viewed from a behavioral decision-making perspective, migrants are not so much choosing to travel a set number of miles, but rather migrants are choosing to travel a set number of miles to reach a specific destination. It is the destination, not the distance, driving migration[2].

Young (1924) found that movement between areas followed an inverse distance squared function and was not linear. Advancements in the 1950s found that parameters tuning the distance value in the form of exponents other than one (Anderson 1956, 1955) produced

[2]In a related paper, Plane (1984) treated distance as a variable to be estimated rather than the number of flows. Plane developed the concept of "migration space" and used flows between origins and destinations to create cartograms warping absolute space to show the relative spatial separation between origins and destination. The high degree of flows between some origins and destinations had the effect of making some origins and destinations appear less distant.

better estimated flows as did parameters modifying the origin and destination population terms (Carrothers 1956; Stewart 1950). In mathematical notation, the spatial interaction model takes the form:

$$T_{ij} = \frac{X_i^{\alpha_i} X_j^{\alpha_j}}{D_{ij}^{\beta}} \tag{3.2}$$

The β, α_i and α_j parameters in equation 3.2 are to be estimated. Developments in the 1970s and 1980s saw further evolution of spatial interaction modelling techniques. Notable advancements are Wilson's (1971) extension of the Newtonian gravity model to include three additional types of models: the production constrained, the attraction constrained, and the production-attraction constrained. These constrains refer to model formulations ensuring that outgoing, incoming, or both incoming and outgoing flows are held fixed to the observed values and enable the determination of individual origin and specific parameter estimates. Researchers demonstrated that including exponents modifying the origin and destination mass terms increased spatial interaction model performance and that the these terms could be empirically determined using a log-normal specification (Curry 1972; Curry et al. 1975; Olsson 1970; Tarver and McLeod 1973) or entropy maximizing technique (Nijkamp 1979; Senior 1979).

Developments in the 1980s saw the identification and inclusion of modeling parameters attenuating the effects of spatial structure (Fotheringham 1983; Fotheringham 1981), a mathematical interpretation of Ravenstein's Laws of Migration (Dorigo and Tobler 1983), alternative formulations of a spatial interaction model (Tobler 1983) and a Poisson regression specification estimating model parameters (Flowerdew and Aitkin 1982). Developments in the 1990s and 2000s featured expansion of the families of spatial interaction models (Pooler 1994), a reengagement with the concept of intervening opportunities (Akwawua and Pooler 2001), and an interest in the concept of spatial choice wherein researchers focus on the decision making progress leading to the selection of a particular destination (Pellegrini and Fotheringham 2002).

Developments in spatial interaction modelling in the latter half of the 20th century focused on the selection of appropriate model coefficients and model calibration techniques. These simultaneous developments were useful for understanding internal migration in relation to post-World War II suburbanization and economic restructuring. This chapter does not feature a new model calibration technique, but rather it engages with a large volume of internal migration data using established modelling techniques enabling a consistent examination of the temporal and spatial trends in 26-years' worth of internal migration data.

The modelling strategy I am using is adapted from Yano et al. (2003) where three covariates - population size, distance, and an accessibility measure - are used to model and compare outgoing flows in Britain and Japan. I will extend this modelling tactic by incorporating an additional variable, a Boolean within-state migration indicator for each county's model. I am choosing to use the modelling tactic and the covariates I have selected for several reasons. The first is that the data are plentiful and that the models are tractable and simple to compute. Second, the coefficients obtained from the subsequent regression models feature intuitive interpretations. Third, the ease of interpretation of the regression coefficients facilitates a succinct and compact understanding of 26-years' worth of internal migration dynamics. Fourth and finally, as a model is prepared for each county in each year in the study period, I can detail the spatialities of migration trends in addition to the amplitude and frequency of migration.

3.2 Model specification

A spatial interaction model generating system-wide parameters takes the form as specified in equation 3.3.

$$T_{ij} = \frac{M_i^{\alpha_i} A_i^{\delta_i} M_j^{\alpha_j} A_j^{\delta_{ij}^{\beta}} S_{ij}^{\gamma}}{D_{ij}} \tag{3.3}$$

Terms specific to the origin are specified with an i subscript and terms specific to the destination are denoted with a j subscript. The mass terms are denoted by M and its

elasticity is denoted by the α exponent. The accessibility measure is indicated by the A term and its elasticity is indicated with the δ exponent. The within-state indicator S applies to the flow between the origin and destination and does not have an origin or destination specific component. Its elasticity is denoted by γ. Likewise, the distance term, D, and its elasticity, β, pertain to the flow between the origin and the destination. In this equation, α_i, δ_i, α_j, δ_j, γ, and β are parameters to be estimated. Equation 3.3 produces parameter estimates for the entire system of migration for the entire US. While interesting and useful (a similar model specification was used in chapter two to aid in the estimation of the aggregate remainder migration flows), this specification does not enable the creation of a migration profile for each county. An origin specific or destination specific model enables the creation of county specific profiles. By estimating a model specific to each origin county's i outgoing flows, the production constrained model becomes:

$$T_{ij} = \frac{M_j^{\alpha_j} A_j^{\delta_j} S_{ij}^{\gamma}}{D_{ij}^{\beta}} \tag{3.4}$$

The α_j, δ_j, γ, and β terms are parameters to be estimated. One could estimate these parameters using a maximum likelihood estimate technique or a regression technique. Both techniques will estimate the parameters for each county which in turn have behavioral interpretations. I will use the regression technique to estimate outgoing migration. The regression framework is enabled by rewriting the above equation to be entirely multiplicative:

$$T_{ij} = M_j^{\alpha_j} A_j^{\delta_j} S_{ij}^{\gamma} D_{ij}^{-\beta} \tag{3.5}$$

And then taking the logarithm of each side and factoring out the exponents.

$$log(T_{ij}) = \alpha log(M_j) + \delta log(A_j) + \gamma log(S_{ij}) - \beta log(D_{ij}) \tag{3.6}$$

This is the log-normal form of the spatial interaction model (Flowerdew 1982). As the migration data are non-negative counts of households moving from one county to another in a set period of time and the movement of households are independent enough from each other

and the population of moving households is large enough, then the number of households moving from i to j will have a Poisson distribution with mean λ_{ij} (Flowerdew and Aitkin 1982). This relationship suggests that the probability of k households moving from i to j is:

$$Pr(k) = \frac{e^{-\lambda_{ij}} \lambda_{ij}^{k}}{k!} \tag{3.7}$$

Equation 3.8 specifies that the probability of k households moving is related to the mean of the Poisson process, λ_{ij}, and the mean of the Poisson process is linked to a linear combination of the logarithmically transformed origin and destination specific variables:

$$\lambda_{ij} = exp(\alpha log(M_j) + \delta log(A_j) + \gamma log(S_{ij}) - \beta log(d_{ij})) \tag{3.8}$$

This assumption enables the fit of a generalized linear model using a Poisson specification (Dennett 2012). This model specification enables a behavioral interpretation of the coefficients because the coefficients are modifying attributes associated with a destination and in aggregate these coefficients pertain to the observed preferences of migrants. This production-constrained model is fit using a regression framework for each county with at least 30 non-zero flows for each year, approximately 77K regression models in total. Each model estimates flows emanating from a specific origin to approximately 3.1K county destinations.

3.3 Migration Data

There are many challenges to working with the IRS county-to-county migration data which necessitate making several assumptions. First, the county-to-county migration data are returns - a proxy for households - moving between counties. These data do not feature any information about individual household composition or socioeconomic status of the household. The lack of household and individual information precludes this study from investigating how gender, income, and family dynamics influence migration (Withers et al. 2008) or how race and education influence the destination selection of migrants (Liaw and Frey 2007, 1998).

This lack of individual detail necessitates a pivot to framing movement in terms of population and the geographic configuration of origin and destination pairs. The spatial interaction models used in this chapter are a way of explaining and describing internal migration trends.

The data driving the analysis in this chapter come from two sources: the IRS's county-to-county migration data and the US Census Bureau's TIGER/Line Files and TIGER/Line Shapefiles for distances between counties. I will first describe the IRS'S county-to-county migration data followed by the US Census Bureau's TIGER/Line Files and TIGER/Line Shapefiles. The IRS's county-to-county migration data supplies the dependent variable and most of the explanatory variables and the TIGER/Line Files and TIGER/Line Shapefiles supplies explanatory variables only.

The IRS's county-to-county migration data are available to the public at no cost and are accessible over the world wide web. The data are available for years 1990 through 2015. As originally downloaded, the data feature the number of returns (a proxy for households), the number of exemptions (a proxy for people), and aggregate gross income moving between pairs of counties. In addition to the county-to-county movement, there are returns, exemptions, and aggregate gross income values for the non-movers in each county. Several examples of previous use of the IRS county-to-county migration data has been to show the growth in western states on account of in-migration (Plane 1999) and the impacts of wealthy migrants on regional metropolitan income distribution (Shumway and Otterstrom 2010). This dissertation contributes to this body of research by looking at internal trends over a longer period, 1990 through 2015, and using enhanced publicly available data.

3.3.1 County-to-county household movement

In each year of IRS migration data, a proportion of the county-to-county returns (and by extension the exemptions and the aggregate gross income) are fully reported - there is a known origin and a known destination - and a proportion of the returns are aggregated into a within-state remainder category and an interstate remainder category. The reason for aggregating the returns is due to disclosure thresholds used by the IRS that protect tax-payers

identity and confidentiality. In each year, a little over 50-percent of all migrating returns feature a fully reported within-state origin and destination and approximately 25-percent of all migrating returns feature an origin in one state and a destination in another. Twenty percent of migrating returns feature an origin in one state and an obfuscated interstate destination, and the remaining five percent of returns feature an origin in one state and an unknown destination within the same state of origin.

Using a combination of spatial interaction modelling and linear optimization programming, I estimated the origins and destinations of an additional 585K county-to-county pairs, on average, per year. Chapter two of this dissertation describes this process in greater detail. This process did not increase the number of migrating households, but it did increase the number of counties with known origins and known destinations by an average of 840-percent per year. By estimating the flows between additional county-to-county pairs, I increased the number of counties for which production and attraction constrained models can be estimated. For example, in the 1999 migration data, there were 548 counties, about 1 in 6, with outgoing migration flows to 30 unique destinations. Including the estimated records brings the total number of counties with at least 30 unique destinations to 3,006, a little more than 95-percent of all counties. In addition to increasing the number of counties suitable for modelling a migration flow, including the estimated county-to-county flows decreases potential biases resulting from omitted flows.

The general trend of internal household migration is that of increasing numbers of migrating households with a rate varying between six percent to seven percent of all households in any given year. Figure 3.1 on page 82 features eight graphics illustrating the dynamics of county-to-county household movement. Graphic C features the legends for graphics A, B, and D through I. Graphic A illustrates the total number of households categorized by interstate and intrastate movement. In all years, slightly more households move within-state boundaries than across state lines. The general trend is that the absolute count of households are migrating compared to the previous year, but there are notable exceptions. In 1990, the number of households migrating within the US was a little more than 6M. The first decrease

in migration was in 1995. In 1996, the number of migrating households continued to rise until 2002 before dipping slightly, likely due to the recession caused by the bursting of the dot-com bubble. The number of households began increasing in 2003 and continued to do so until 2006, an increase is visible in 2007, followed by decreases in 2008 and 2009 due to the Great Recession. By 2010 the number of households migrating was again increasing. Moves that were postponed due to the Great Recession were finally realized. The sizable drop in 2014 is due to differences in data processing techniques. In 2015, more households were again on the move. The IRS data do feature several key differences in migration when compared to other federal data sources. Notably, the IRS migration rates come from the universe of tax filers and is not a survey of the population like the Decennial Census, the American Community Survey, or the Current Population Survey. Some households do not file taxes such as the elderly and those with low income (Gross 2009). In addition, according to the Current Population Survey, 87-percent of households filed taxes between 1992 and 2009 and that tax filers migration more frequently than households that do file (Molloy et al. 2011, p. 178)

Graphic B features county-to-county movement categorized by the metropolitan and non-metropolitan classifications of the origins and destinations. The largest share of moves are households moving from one metropolitan area to another followed by non-metropolitan-to-metropolitan movement. The large share of metropolitan-to-metropolitan movement is to be expected given that metropolitan areas have large population bases and are therefore generating large attractivity and emissivity quotients. What is less intuitive is the difference in the non-metropolitan-to-metropolitan movement and metropolitan-to-non-metropolitan movement. The non-metropolitan-to-metropolitan movement reflects continued net outmigration from non-metropolitan counties (Johnson and Lichter 2019) and the metropolitan-to-non-metropolitan movement reflects amenity and retirement migration (Cadieux and Hurley 2011; Nelson and Nelson 2011). Graphics D through I showcase the counts of emigrating households by each county for years 1990, 1995, 2000, 2010, and 2015. These maps show how numbers of out-migrating households align with more populated counties. Metropolitan

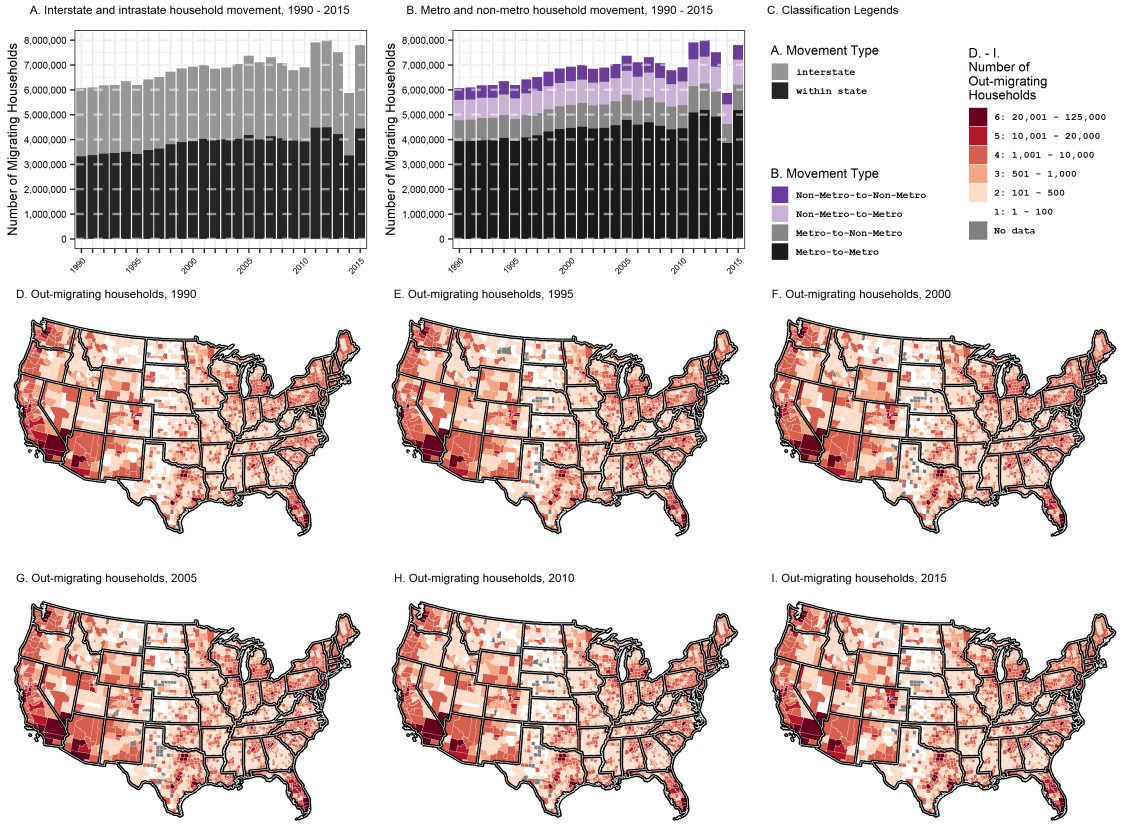


Figure 3.1: Statistical and spatial distributions of county-to-county movement, 1990-2015

counties on the west coast and east coast send more households while counties in the central plains states send fewer households.

Figure 3.1 is the first example of the visualization technique used throughout this chapter: a large visualization consisting of multiple smaller graphics such as boxplots, histograms, bar plots, maps, and line plots (Tufte et al. 1990). This is done to maximize the display of the volume of information generated from these data, enable year-over-year comparison, and minimize the amount of space consumed by any one graphic in this chapter (and dissertation).

3.3.2 On the decision to model outgoing flows

One unique aspect of the IRS's county-to-county data is that migration flows are recorded for both households entering a county and households leaving a county. Within the context of this data, the origins and destinations of movement are fully recorded. Because of this, it is possible to fit unconstrained, singly constrained, or double-constrained spatial interaction models (Wilson 1971) along with origin specific and destination specific variants. The combinations of constraints and origin and destination specificity enable different ways of modeling incoming migration and outgoing migration and each strategy has advantages and disadvantages. As this chapter is about the spatiality of migration during a 26-year period a case could be made to examine the spatiality of migration by focusing on either direction of flows. However, I will examine the spatiality of migration by focusing on outgoing migration because of trends observed in the data and the interpretation of the coefficients generated from the spatial interaction models.

The two trends in the data I wish to focus on to make the case for studying outgoing migration is a unique way of measuring the degree to which counties exchange migrants - the cosine similarity - and the net migration rate. Combining these two measures provides a comprehensive and succinct overview of the US's yearly migration profile. As of the writing of this dissertation, I am unaware of any migration studies making use of the cosine similarity. The cosine similarity, however, is used in natural language processing and artificial intelligence research (Sidorov et al. 2014). The cosine similarity measures the similarity in

direction between two vectors. The cosine similarity only considers similarity in direction and not magnitude. A vector of [0,1] has the same direction as a vector of [0,2] even though the magnitudes of the two vectors are different. Each county's incoming migration values and outgoing migrations can be viewed as acting in approximately 3,100 dimensions with the number of migrants acting as the coordinate in a dimension. In this sense, the cosine similarity measures the degree to which a county is exchanging migrants with other counties. For every county x there is the set of possible origins A and the set of possible destinations B. These sets are identical, but specific to each county. A number of migrants enter focal county x from each origin A_k and a number of the migrants depart focal county x to each destination B_k. The more similar the values are between the number of migrants entering county x from each origin A_k and departing from county x to each destination B_k the higher the cosine similarity score. The cosine similarity for county x is defined as:

$$Degree\ of\ Migrant\ Exchange_x = cosine\ similarity_x = \frac{\sum_{k=1}^{n} A_k B_k}{\sqrt{\sum_{k=1}^{n} A_k^2}\sqrt{\sum_{k=1}^{n} B_k^2}} * 100 \quad (3.9)$$

Rather than refer to this value as the cosine similarity, I will refer to it as the degree of migrant exchange or DME for short. The DME ranges from 0 to 100 and higher values indicate a greater degree of exchange[3]. While the majority of DME values are far more often closer to 100 than not, values closer to 0 are indicative of counties with orthogonal migration streams. A value of 0 indicates that a county is sending and receiving migrants from completely different sets of destinations and origins: there is no overlap in the set of origins and destinations. Counties with a DME value close to 0 can be seen as redistributive counties and pivots. This measure extends previous work on population redistribution as seen in (Plane and Mulligan 1997; Rogers and Raymer 1998; Roseman and McHugh 1982). While previous studies have used the Gini-index or the migration effectiveness ratio to prepare

[3]A DME less than zero is not possible because migration of households is always a non-negative count. In other applications however, it is possible to produce a negative cosine similarity value indicating a completely opposite relationship between the input vectors.

a value reflecting migration in a system as whole, the DME is a term calculated *for each subunit* in the migration system.

I computed the DME for each county that has both outgoing and incoming flows in each year. I then computed the net migration rate (NMR) for each county (Newbold 2010). The NMR per 1,000 households of county k is defined as the number of incoming households I_k minus the number of outgoing households O_k divided by the number of non-migrating households P_k multiplied by 1,000. This value is expressed as:

$$NMR_k = \frac{(I_k - O_k)}{P_k} * 1000 \tag{3.10}$$

Negative values indicate that more migrants are leaving an area than entering and positive values indicate that more migrants are entering an area than leaving. Theoretically, the NMR has no lower bound and no upper bound, the NMR is limited by the size of the non-migrating households. In the enhanced IRS migration data, 98-percent of the combined, all-years distribution of the NMR is between the values of (-45.7, 53.1) per 1000 households. The median value is -1.8 and the average value is 0.1. The negative median indicates that over time households are concentrating in fewer destinations. I classify counties as having a neutral net migration rate as counties with an NMR between the open interval (-1, 1). Irrespective of a county having a positive, negative, or neutral net migration rate, most counties feature high DME values meaning that for a given county, most migrants are coming from and going to the same set of origins and destinations. For example, in 1999, 46,614 households migrated to King County, Washington from 1,084 origins and 50,434 households migrated to 1,256 destinations from King County while 654,744 households were not mobile. This works out to an NMR of -5.83 households and a DME of 0.98. The high DME value is indicative of a high degree of exchange. There are only 594 counties in both the set of origin counties and the set of destination counties for King County migration in 1999. However, these 594 counties account for 94-percent and 95-percent of the incoming and outgoing migrants, respectively. Indeed, gravitational theories of movement suggest values such as these as the degree of exchange between sets of origins and destinations is predicated

on the origin and destination masses and spatial separation.

Figure 3.2 on page 88 features multiple graphics illustrating the statistical and spatial distributions of the degree of migrant exchange and net migration rate measures. The box-plots in graphic A show the distribution of the DME for all counties in each year, regardless of the net migration rate. Graphic B features an additional net migration rate classification scheme: negative $(-\infty, -1]$, neutral $(-1, 1)$, and positive $[1, \infty$. Graphic C features a legend pertaining to the bar plot in graphic B and the maps in graphics D through I. Graphic A shows that inter-quartile range of the DME is shrinking indicating that migration is happening to and from established destinations and origins and those origins and destinations are swapping migrants to a high degree. In 2013, the interquartile range starts to expand indicating that origins and destinations are swapping migrants to less of a degree. Graphic B classifies a county as having a positive, negative, or neutral NMR along with a low, medium, or high DME and counts the number of counties in each of the combined NMR-DME categories over time. The trend is that, apart from the first half of the 1990s, there were more counties with a negative NMR than with a positive NMR. This means that more and more counties were urbanizing and more households were moving to metropolitan counties (see Figure 2.14 on page 69 for graphics showcasing the increase in non-metropolitan-to-metropolitan migration). The high DME values of the counties combined with a negative NMR add an additional facet to the interpretation of the NMR values. A negative NMR simply means that more people departed a county than entered and the counties with a high DME and a negative NMR suggest that there is reciprocal negative migration meaning that people are still moving to areas with a negative NMR. Rural migration trends, and urban to rural amenity migration in particular, are discussed at length in Nelson and Nelson (2011) and Nelson et al. (2009).

Graphic B in Figure 3.2 features a stacked bar plot of the same data as in graphic A with an additional classification scheme. Each county is classified into a positive, neutral, and negative NMR and a low (0-75-percent), medium (75-95-percent), and high (95-100-percent) DME. The DME categories were chosen based on the distributions of the DME values. The

trend over time is that there is a growing number of counties with a negative NMR. That is, there are more counties with more out-migration than in-migration and larger numbers of migrants increasingly going to fewer destinations suggesting that population is concentrating. With respect to the increasing numbers of counties with high DME values, irrespective of a positive, neutral, and negative NMR, that is partly explained by the fact that most moves are of shorter distance. Graphic B in Figure 2.13 on page 67 features the share of movement by county adjacency. In any year, approximately 35-percent of all moves are to a directly adjacent county. Because of the geometric structure of county boundaries, there are only so many directly adjacent counties available to move to resulting in the across-the-board high DME values.

Graphics D through I show the spatial distribution of the DMR-NMR categorization for years 1990, 1995, 2000, 2005, 2010, and 2015. These maps show the positive net migration rates in the western portions of the country. Counties in California frequently feature negative net migration rates (the California exodus (Henrie and Plane 2008)), while counties in other western states (Washington, Oregon, Idaho, and Arizona) show positive trends. Negative net migration rates can be seen in southern and rust belt counties. The increase in counties with negative net migration rates in 2010 is a response to the Great Recession.

While a good number of counties feature high DME values, plenty do not. One could imagine that counties with a high DME and a positive NMR and counties with a negative NMR and a high DME are counties with a history of sending and receiving. That is, these counties are paired. The counties with mid-to-low DME values suggest several things, depending on the sign of the NMR. For counties with a positive NMR, a mid-to-low DME indicates an increase in urbanization. Counties with a neutral NMR and a mid-to-low DME are frequently outer metropolitan counties nearby large population centers. Finally, counties with a negative NMR and a mid to low DME are indicative of counties with migrants selecting destinations other than counties that supply migrants. Potential reasons for this could be life-course migration such as students and urban-to-rural migration for retirement or a response to an economic shock in a region such that movement to nearby counties is

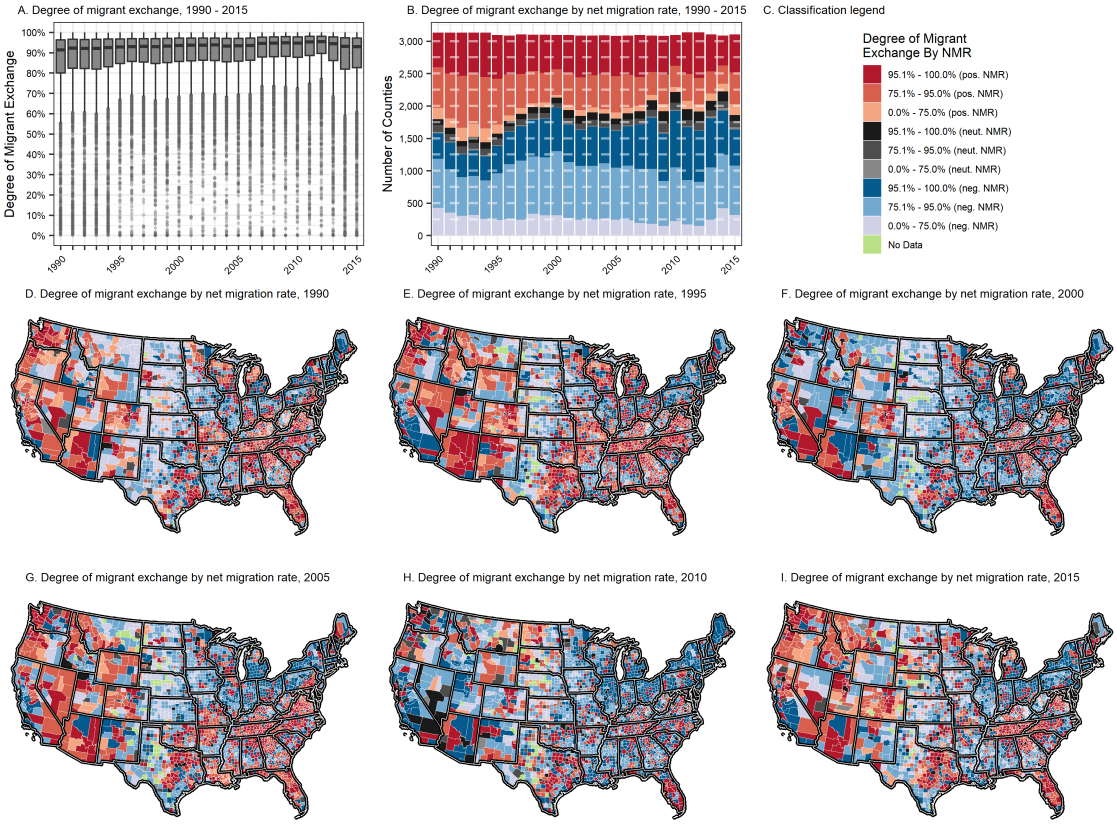


Figure 3.2: Statistical and spatial distributions of the degree of migrant exchange by net migration rate, 1990-2015

still too near the shock.

I will focus on the outgoing flows in this chapter for three reasons. First, as shown in Figure 3.2, there are more counties with a negative NMR meaning that more households are migrating to fewer destinations. The spatial interaction models in this chapter will shed insight on those preferences. Second, even with counties featuring a positive NMR, there is a still an outbound flow. Third and finally, the coefficients from the spatial interaction models feature a behavioral interpretation showcasing the destination preferences of households in each county.

3.3.3 Destination mass

The destination mass variable is operationalized as the number of non-migrating returns in any county each year. These values are appropriate to use for destination masses as the non-migrating returns are a proxy for non-migrating households in the same way that the migrating returns function as a proxy for migrating households. As tax returns measure occupied households, it is not possible to perform a direct comparison of non-migrating occupied households with non-migrating returns for every year in the 1990 through 2015 period. However, it is possible to perform a comparison with all occupied households and all tax returns in Decennial Census Years 1990, 2000, and 2010. While the decennial censuses do track migration, they only track migrating individuals and not households. I can benchmark the US Census Bureau's county population estimates with the number of exemptions per county for each year.

Figure 3.3 on page 90 features the counts of occupied households and people juxtaposed against the counts of returns and exemptions by data source. Data originating from the US Census are shaded grey and data originating from the IRS are shaded purple. Graphic A features a bar plot comparing the 1990, 2000, and 2010 decennial census counts of occupied houses with the 1990, 2000, and 2010 total counts of non-movers and movers combined). In each decennial year, approximately 97-percent of occupied households file a tax return. Graphic B features the distribution of occupied houses per county and the number of total

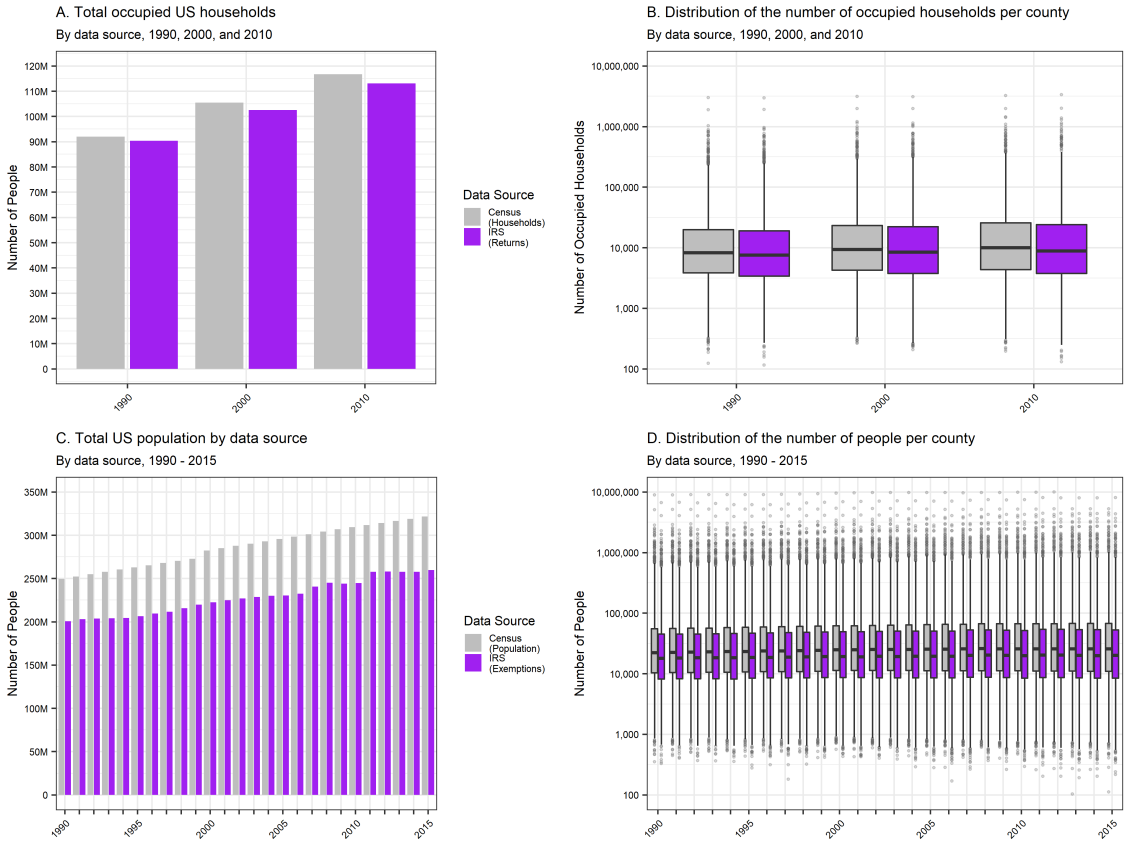


Figure 3.3: Comparison of occupied households and total population by data source, 1990-2015

returns per county for years 1990, 2000, and 2010. The shape and general extents of the distributions match. Graphic C features the total number of people in the US by data source for years 1990 through 2015. The Census Bureau reports the number of people and the IRS reports the number of exemptions. In general, the number of exemptions in any given year is approximately 80-percent of the US population due to not all households filling a tax return (the elderly and those not meeting a certain income thresholds). Graphic D features the same data as graphic C, only expressed on a per-county basis. In general, the shape and extents of the distributions match.

Figure 3.4 on page 91 focuses strictly on counts non-moving households (as opposed to the

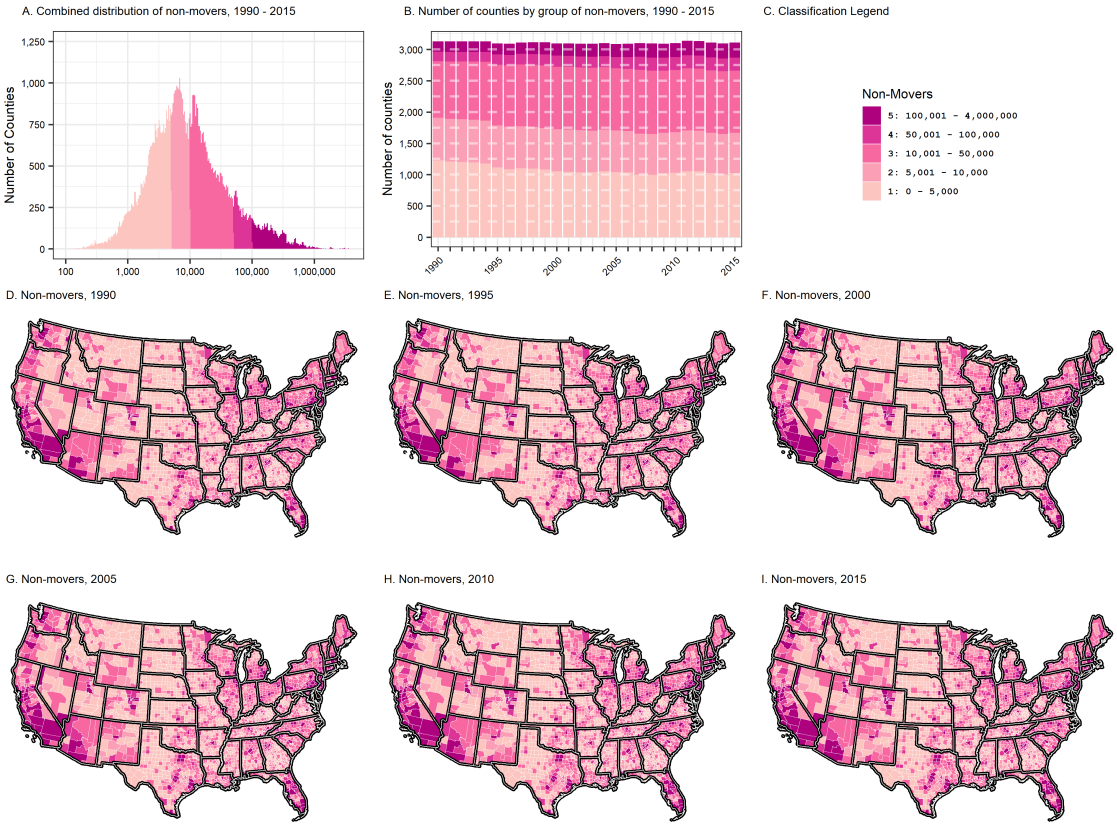


Figure 3.4: Comparison of occupied households and total population by data source, 1990-2015

total households as featured in Figure 3.3). Briefly, there is a slight decrease in the number of counties with less than 10,000 non-moving households and an increase in counties with more than 10,000 non-moving households. The spatial distribution of counties with more than 50,000 households follows metropolitan counties. The maps for years 1990, 2000, 2005, 2010, 2015 show counties with increasing numbers of non-movers in Texas, New Mexico, Arizona, California, Oregon, and Washington reflecting trends in the growth of western states. In sum, the number of non-migrating returns in a county corresponds with the number of occupied households in a county and follows recognizable settlement patterns and will therefore be used as the destination mass in the spatial interaction models.

3.3.4 Distance

The data behind the county-to-county distance measurements comes from the US Census Bureau's TIGER/Line Files and the TIGER/Line Shapefiles, referred to as TIGER/Files for convenience. The TIGER/Files are used for two reasons. First, the TIGER/Files function as each year's manifest of existent counties. I can benchmark the counties in the IRS county-to-county migration data to the TIGER/Files to understand which counties feature migration and which do not as the IRS only reports counties that feature some form of migration. Second, the TIGER/Files feature accurate and precise county boundaries which is necessary for creating accurate and precise county-to-county distance measurements. Simply put, county boundaries did not remain constant during the 1990-2015 period. County boundary changes since 1970 are documented by the US Census Bureau (US Census Bureau 2019). Effort was made to account for the shifts in county annexations, county incorporation, and county disincorporation and the 19 different vintages of TIGER/Files and reflects the then-most recent changes. Incorporating all 19 different vintages of the TIGER/Files enables the construction of accurate county-to-county distances and county adjacency matrices. The spatial separation between each pair of counties is measured as the number of miles between the centroid of each origin county and the centroid of each destination county. E, page 231, features a more in-depth description of the process used to extract the county geometry, harmonize the boundaries, create centroids for each county geometry, and calculate the centroid-to-centroid distance.

3.3.5 Accessibility

The accessibility measure used in the modelling of the flows in this chapter was first described by Fotheringham (1983, 1986). The measure is meant to address spatial structure and miss specified distance decay parameters (Fotheringham 1981). Spatial structure is defined as the size and configuration of origins and destination within a spatial system. Fotheringham's (1981) article wrestles with the issue that the distance decay parameter is in part a function

of spatial structure and failure to include measures of spatial structure in a spatial interaction model produces incorrect parameter estimates (Fotheringham and Webber 1980; Curry et al. 1975; Johnston 1973). The accessibility measure is defined mathematically as:

$$A_j = \sum_{k(k \neq j) \frac{P_k}{D_{ij}}} \tag{3.11}$$

The accessibility of area j is defined as the sum of the population of area k for all $k \neq j$ divided by the distance between j and k. This measure captures both the accessibility and the degree of competition a location faces from other, more nearby locations. Higher values of A_j indicate that j is more accessible and therefore competes to a greater degree with other nearby destinations. Lower values of A_j suggest that j is less accessible and does not compete with other nearby destinations.

I computed the accessibility measure for each county for each year using the number of non-moving returns in each county and I have included a nine-part graphic featuring the statistical and spatial distributions of the accessibility measure. Eight of the nine graphics in Figure 3.5 on page 94 visualize a component of the distribution of the accessibility measure while a legend pertaining to graphic B and graphics D through I. In all graphics, higher values indicate a county is more accessible to the population of other counties. The accessibility index, for visualization purposes, has been divided by a factor for 10,000 to help with visualization. The index ranges from two to more than seventy.

Graphic A Figure 3.5 features a boxplot for each year showing the general trend in the accessibility index over time. As the distance between counties has remained effectively constant over time and the population of the US has increased over time, the accessibility index has generally increased for each county over time, as indicated by the upward trends of the 25th-percentile, median, and 50th-percentile. Graphic B features a histogram of the pooled county accessibility measures and the histogram is shaded according to quintile values. Blue shades indicate less accessible counties and red shades indicate more accessible counties. The small cluster of values near the bottom of the distribution represent remote counties.

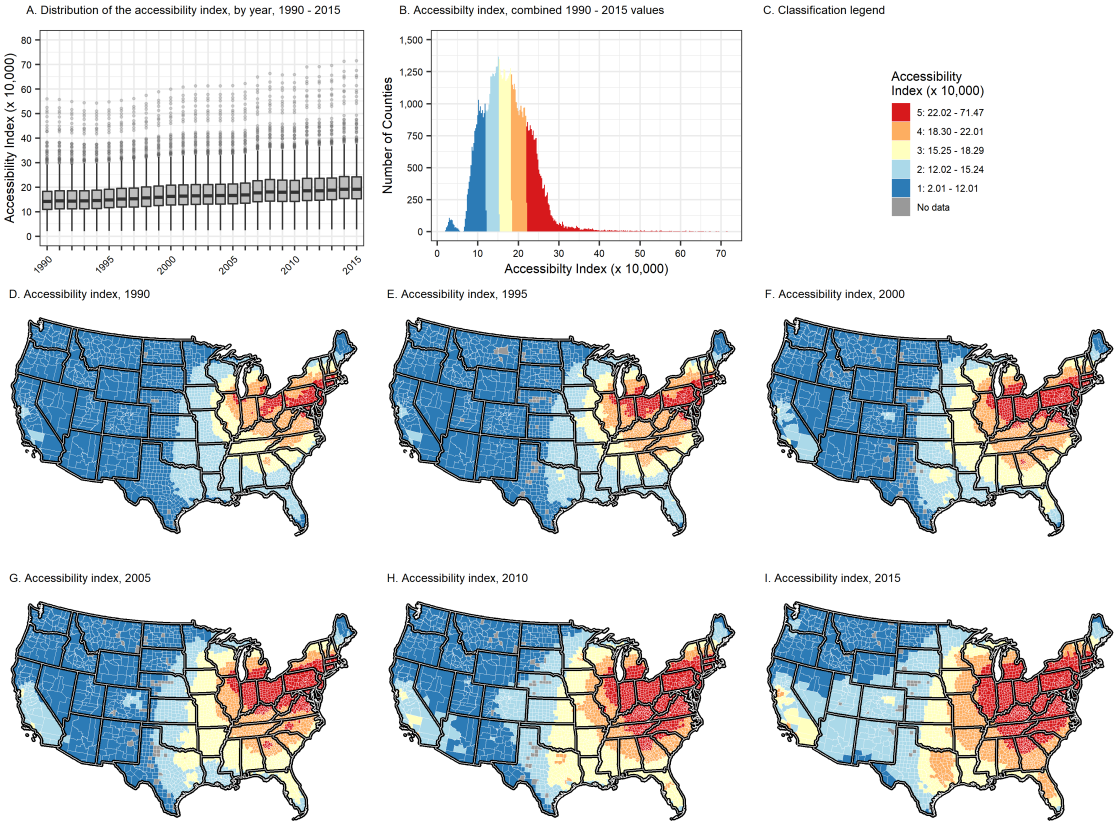


Figure 3.5: Statistical and spatial distributions of the accessibility index, 1990-2015

Graphics C through H are maps depicting the accessibility measure for years 1990, 1995, 2000, 2005, 2010, and 2015. Each county in each map is colored using the same coloring scheme as in graphic B, the 26-year period all county pooled accessibility index values. In graphic C, the map pertaining to 1990, the most accessible counties are in Ohio, Pennsylvania, New Jersey, and New York. This is on account of counties being smaller in areal dimension and the population being more numerous in this area of the country. Accessibility decreases as distance from the rust belt increases. Accessibility is not a metropolitan measure or a rural measure; it is a measure relating the proximity of the population of every other county to a focal county. Therefore, Cook County, Illinois has a lower accessibility value than its immediate neighbor DuPage County, Illinois. The population of Cook County, Illinois is considered in the calculation of the measure for DuPage County, but the population of Cook County is not considered in the calculation of the measure for Cook County. In a similar fashion, Lake County, Illinois, Will County Illinois, and Lake County Indiana all feature higher accessibility measures than Cook County. In practice, this means that these counties compete more to attract migrants from other regions.

Over the course of the 1990 through 2015 period, the concentration of the accessibility measure deepens in the rust belt and emanates further north and further south along the eastern seaboard and into Kentucky, North Carolina, Georgia (especially around the Atlanta metropolitan area) indicating that in each year each county is becoming more accessible to the population of other counties. Beginning in 1990, nearly all counties in western states are generally less accessible. This is due to generally lower population density in the western portion of the country. In 2000, counties around the Dallas-Ft. Worth Metro in Texas, the Bay Area in California, and Southern California exhibit growth in accessibility. By 2010 and 2015, the increase in accessibility is much more noticeable. This is due to population growth in the country as a whole and population growth in the aforementioned areas.

A final note on the accessibility variable. Fotheringham (1983) describes how the inclusion of the accessibility variable helps alleviate the misspecification of the distance parameter. But what about model accuracy, what does the inclusion of this variable do for model per-

formance? I created a scenario using the 1999 data and prepared two production constrained models for each county. The first model featured the accessibility measure and the second model did not. Approximately 6.1K models were run to generate these test data. Upon the completion of each model, I computed the root-mean-square error (RMSE) for each model. In 64-percent of the models, including the accessibility measure improved model fit by reducing the RMSE by an average of eight percent. In the other 36-percent of the models, including the accessibility measure increased the RMSE by an average of a little more than two percent, indicate a poorer fit. From both a theoretical and empirical perspective, I will include the accessibility measure.

3.3.6 Within-state migration

The within-state variable is a Boolean variable indicating whether movement originates and terminates in the same state of origin or not. While the accessibility measure controls for the effect of spatial structure via how county population is apportioned over the spatial separation between counties, the within-state migration measure controls for the explicit decision to migrate within the state of origin or move out of the state of origin. Kone et al. (2018) use data from the 2001 Census from India illustrating how internal state borders are barriers to migration. In addition, Griffith (1982) found that the geometric shape of origins and destinations within a region does influence spatial interaction within that region. In other words, there is a scalar component to spatial interaction that I am accounting for with the use of a within-state migration Boolean variable. In addition, the within-state variable helps control for long-distance in-state moves (a household moving to west Texas from east Texas, for example). Like the accessibility measure, it is important to determine empirically the usefulness of including this variable. Using the 1999 data I created a similar testing scenario with one set of production constrained models featuring the additional within-state Boolean indicator and another set of production constrained models without the within-state Boolean indicator variable. I fit 6.1K models in total. The results indicate that approximately 50-percent of the time, the model with the within-state variable outperforms the model without

the within-state variable, as judged by the RMSE value. When including the within-state model results in a better fit, the fit is improved by 15-percent on average. When the inclusion of within-state model results in a poorer fit, the fit is worsened by 7-percent on average.

I will include the within-state variable based on these diagnoses and what will be gained theoretically from its inclusion: an insight into the preferences for interstate migration. Finally, while I have discussed county adjacency in relation to the geometric structure of counties, I am not including a county adjacency variable because it is highly correlated with the within-state variable: 86-percent of the approximately 19K county adjacency pairs are within the same state.

3.4 Model discussion

The coefficients from the spatial interaction models are obtained from a linear combination of logarithmically transformed variables pertaining to mass, accessibility, distance, and a within-state migration indicator. This section begins with a brief discussion on model fit. Each of the four coefficients are described in detail after several remarks on the initial interpretation of the coefficients and visualization strategies used.

The percent of deviance explained is used to assess model fit for each county, for each year. Higher values indicate a better fit. Figure 3.6 on page 98 features boxplots for each year of the distribution of the percent of the deviance explained for each county. Boxplot visualizations do not include the average value of a distribution and so the red dots were added to indicate the mean value of the percent of the deviance explained for each year. The median ranges from 70-percent in 1990 to 71-percent in 2015 with a peak of 75-percent in 2011. The mean follows a similar trajectory. Beginning in 2011, the outliers became more dispersed indicating that the model is not performing as well in some counties. I believe this is reflecting a combination of the change in the IRS'S data tabulation processes and changes and less migration trends (see DeWaard et al. (2020b) and Pierce (2015) for a discussion of the changes to data processing).

For each regression coefficient, only values within the range of 0.01-percent to 99.99-

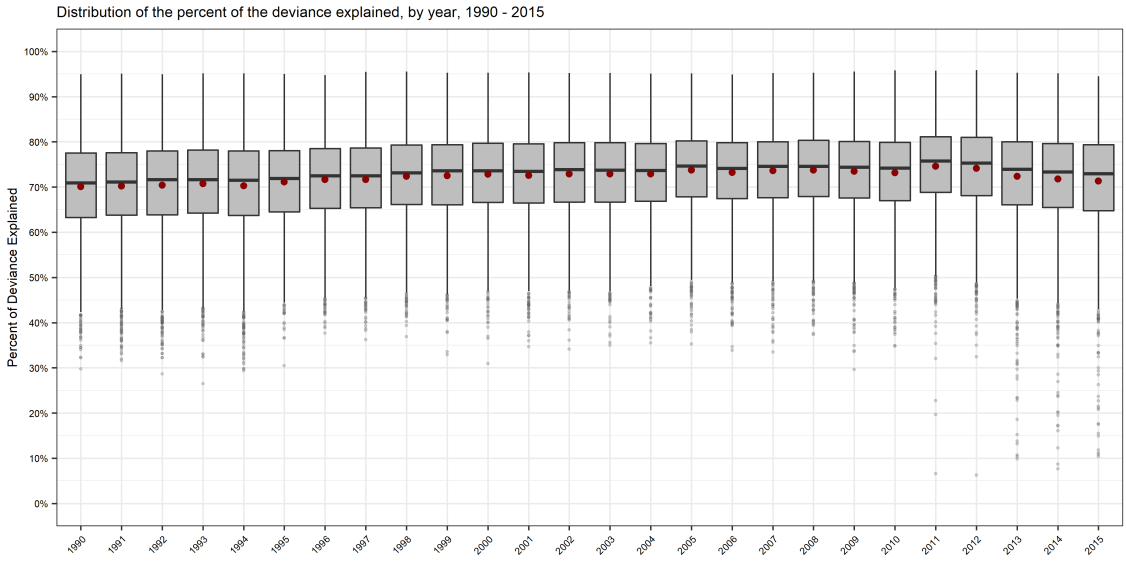


Figure 3.6: Distribution of the percent of the deviance explained, 1990-2015

percent of the distribution are used, inclusive across all years. For each variable originating from the spatial interaction models, I will include two visualizations and each visualization will have multiple smaller graphics. The first visualization in the set of two visualizations features four graphics. Graphic A features two boxplots per year: a boxplot with coefficient values and a boxplot for the significant coefficients only. These boxplots showcase the distribution of a specific coefficient by year. Graphic B features a histogram of the pooled all-years coefficients. The histogram is divided into five groups and each group is colored distinctly using a diverging color ramp. In addition, a second histogram featuring the outlines of only the significant coefficients is placed over the first histogram showcasing the distribution of significant-only coefficients at the 0.05-level. A legend is included in graphic B highlighting the cut points for each group. Graphic C in the first visualization features the counts of the number of counties in each group in each year and graphic D expresses those counts as a percentage showcasing the relative year-over-year change. The same colors used in graphic B are used in graphics C and D. Each coefficient's groups are static over the 26-year period enabling the comparison of group change over time. Shading by group illustrates the degree

to which the size of the groups change over time.

The second visualization features six maps, a histogram, a line plot, and a legend. The six maps, graphics A through F, feature the spatial distribution of the coefficients for the years 1990, 1995, 2000, 2005, 2010, and 2015. Not that models were produced for each county with at least 30 destinations. This often included counties in Alaska and Hawaii. However, due to space constraints, the maps omit Alaska and Hawaii. For the six maps, cross-hatching is applied to counties with non-significant coefficients. This is done to illustrate the spatiality of the non-significant coefficients. Graphic G is the same histogram as featured in graphic B in the first visualization and is included for reference. Graphic H is a line plot derived from the data generating the bar plot in graphic D in the first visualization and showcases the percent of the counties in each group per year. This line plot illustrates the dynamics of the groups over time. Included in the second visualization is a legend featuring the group cut points and a "no data" color for counties without a coefficient. For graphics B, C, and D in the first visualization and all graphics in the second visualization, the same group categorization and coloring schemes are used enabling a consistent comparison across graphics and visualizations. Each color ramp features five colors and is assigned a value of one through five. Bin ranges were chosen to highlight differences within each coefficient's distribution. The third color in each color ramp is a shade of grey and includes the mean and median of each coefficient. Four distinct color ramps and the associated cut points are on display in Figure 3.7.

The color ramps, and other colors throughout this dissertation, are inspired by Professor Cynthia Brewer (2005) and are used in the visualizations of these data. While the coefficient data are continuous and most information design specialists recommend a color ramp ranging from a light hue to a dark hue for continuous data (Ware 2019), I am foregoing the tradition in favor of diverging color schemes to highlight differences in the distributions of these coefficients. When combined with a five-category grouping scheme, a diverging color ramp tracks neatly with each group and the different shades indicate the statistical and spatial relationship of a value to the entire coefficient distribution.

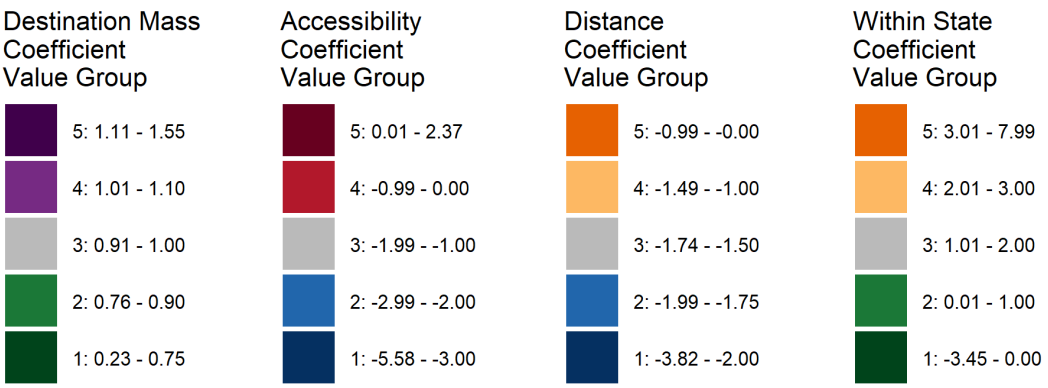


Figure 3.7: Coefficient color ramps and coefficient group values

A common theme throughout all the visuals presented in this chapter is the noticeable differences in years before 2013 and years 2013 and onwards. My investigations and analyses have led me to conclude that these differences do not solely reflect changes in migration. Rather, some of these changes are the result of changes in the data tabulation processes used by the Census Bureau and the IRS. For example, beginning in 2010, the IRS took control of preparing the migration data while it had previously been prepared by the Census Bureau (Pierce 2015). In addition, for years 1990 through 2012, the minimum number of flows between counties to warrant a disclosure was 10. Starting in 2013, the minimum number increased to 20 (IRS 2016). Finally, based on a correspondence between the Missouri Census Data Center and a representative from the IRS's Statistics of Income Branch (Pierce 2020), changes in the data pre-2013 and 2013 and later data are attributed to an increase in the IRS'S effort to combat identity theft.

3.4.1 Initial interpretation

The use of a regression framework permits a statistical interpretation and a behavioral interpretation of each regression coefficient. The statistical interpretation is the same for all four coefficients. A coefficient less than zero indicates that a facet of a destination is not seen as desirable by the migrants. This has the effect of decreasing the flows emanating from a county. Coefficients greater than zero and less than one dampen the expected flow of households out of a county, but still contribute a positive component to an outbound flow. This means that an aspect of a destination for migrants in a county is discounted, but not thought of as a negative. Coefficients greater than one amplify the expected flows out of a county meaning that migrants see a destination's characteristics as positive and this destination characteristic serves to amplify flows. I will focus on the behavioral interpretation of each coefficient in each section.

3.4.2 Destination Mass

Each of the four graphics in Figure 3.8 showcase a different facet of the distribution of the destination mass coefficient. Graphic A features two boxplots for each year in the distribution of the coefficient. The destination mass coefficients range from 0.23 to 1.55. The light grey boxplot features all coefficients and the dark grey boxplot showcases only the significant coefficients at the 0.05-level. Of the 77K accessibility coefficients, over 99-percent are significant at the 0.05 level. The range, median, and interquartile range of all coefficients and only the significant coefficients is consistent over time with slight changes around 2000 and in the years 2013 through 2015. Graphic A indicates that nearly all coefficients are significant at the 95-percent level. This is a trend that is common in both the population mass coefficients and the distance coefficients but less so with the accessibility and within-state migration coefficients. A coefficient less than 1.0 indicates that households are moving to counties with relatively fewer households and values greater than 1.0 indicate that households are moving to counties with relatively more households than the focal county. The mean and median of

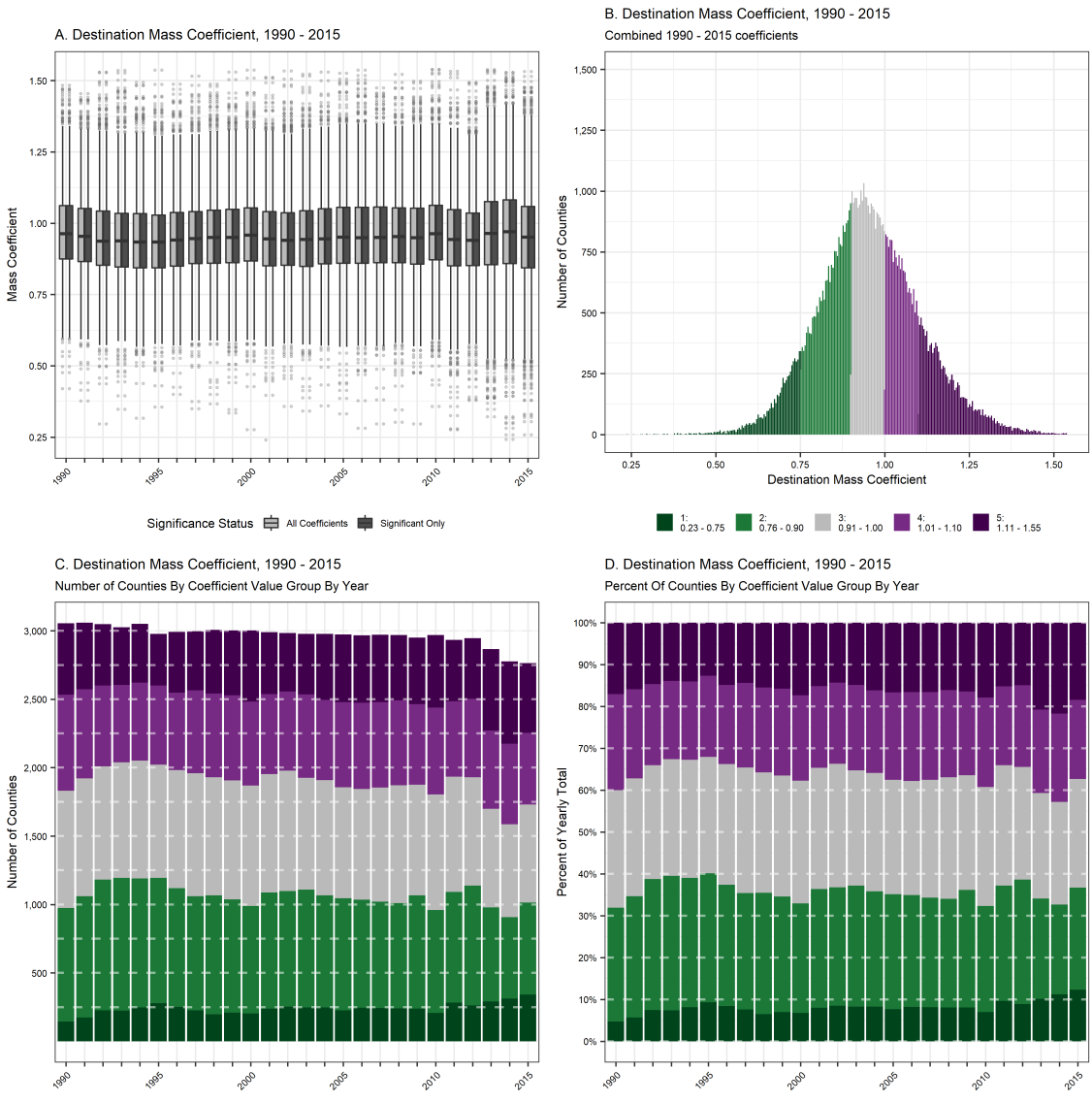


Figure 3.8: Statistical distributions of the destination mass coefficient, 1990-2015

the destination mass coefficient are between 0.91 and 1.00 for each year indicating that most moves are to counties of similar size or slightly smaller. Graphic B features a histogram of the pooled coefficients for all years with each group shaded a color from purple to green. The green values indicate a preference for less populated destinations and the grey values indicate a preference for similarly populated places. Purple values indicate a preference for relatively more populated areas.

Graphic C in Figure 3.8 shows the counts of the number of counties in each group in each year and graphic D expresses those counts as a yearly percentage, to showcase the proportion of change and normalize by the counts of counties in any given year. If the number of counties in each group were to remain constant over time, the colored bars would be horizontal. In this case, there is growth in the number of counties in groups one through five indicating that over time, there has been a growth in the preference for relatively less populated areas. This does not mean a growth in urban to rural migration per se, but rather a growth in the migration of households from a more populated county to a less populated county, a trend previously documented by Plane et al. (2005).

Figure 3.9 features six maps, a histogram, a line plot, and a legend. In each map, the general trend is that households leaving rural areas in western states prefer relatively less populated destinations and households leaving metropolitan areas in western states prefer relatively more populated areas. The trend is reversed for areas east of the Mississippi river. Over time, Montana, North Dakota, South Dakota, Nebraska, Kansas, Oklahoma, and Texas feature more and more counties with no data. This is indicative of these counties not generating migration flows. In aggregate, it appears that households are showcasing an increase in preference for less populated areas and more populated areas as exhibited by the growth of groups 1 and 5 in graphic H in Figure 3.9. This is the simultaneous trend of moving to metropolitan areas and moving from a larger metropolitan area (group 5) and moving to a smaller metropolitan area (group 1).

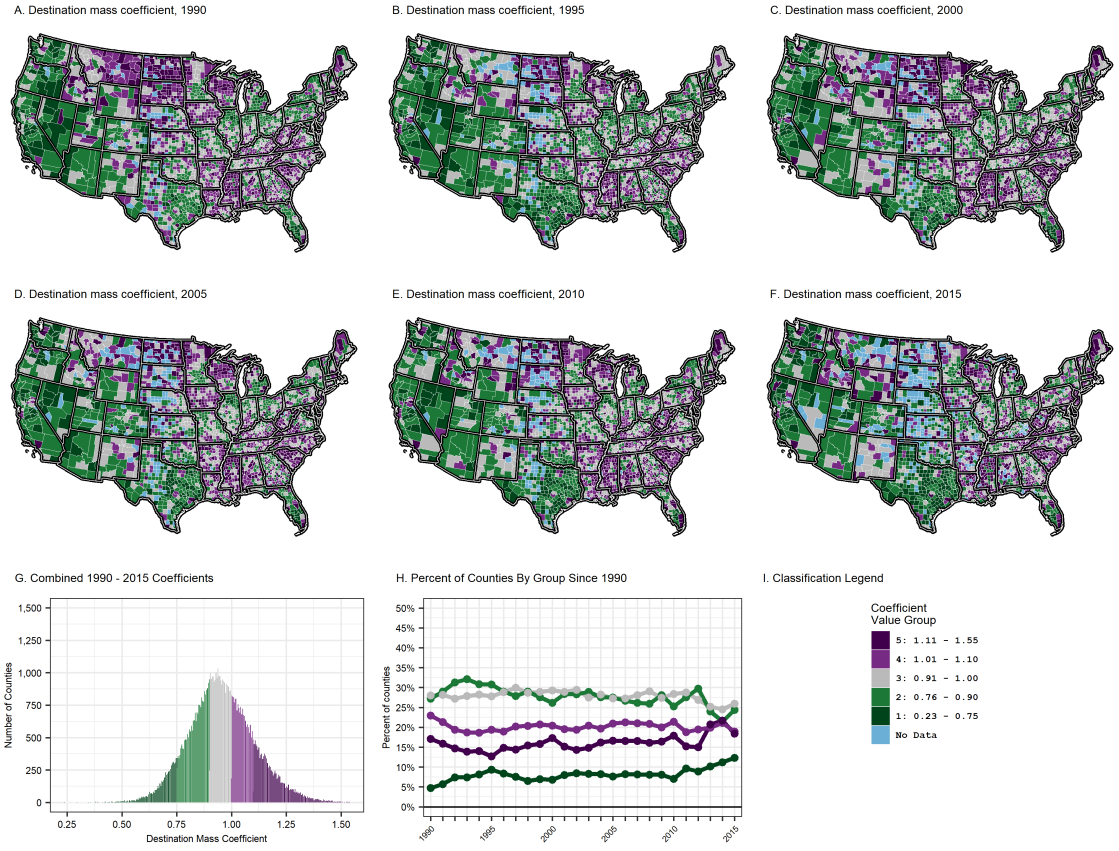


Figure 3.9: Spatial distributions of the destination mass coefficient, 1990-2015

3.4.3 Distance

While the population of the US has increased, the distance between counties has remained fixed. Accordingly, the distance decay coefficient should become less negative as gravitational theories of movement predict greater interaction between the origin and destination as the population of the origin and destination increase. And this would be true if population growth were uniform across the US. As most moves are of shorter distance and with population concentrating in metros, we see fewer households moving greater distances. Values of the distance decay component closer to zero are indicative of a population preferring longer distance moves while more negative distance decay coefficient values are indicative of a population preferring shorter distance moves. Because these coefficients were created for each county for each year, the interpretation of the coefficients is relative: the coefficient of distance is the rate at which interaction (migration) decreases with distance. More negative values indicate that interaction decreases with distance more quickly.

The coefficients in Figure 3.10 range from -3.82 to 0.0. Graphic A shows we see that there is near agreement in the distributions of all coefficients and the significant coefficients at the 0.05-level only. Of the 77K distance decay coefficients, over 99-percent are significant at the 0.05 level. The median and inter quartile ranges are decreasing from 1990 through 2009 indicating that more households were moving shorter distances. There is a slight uptick in the median and interquartile range in 2010 followed by a decrease in 2011 and 2012. In 2013, both the median and interquartile ranges increased indicating that more households were moving longer distances, due to recovery after the Great Recession. These trends are even more visible in graphics C and D. In 1990 through 2009, there is growth in the number of counties in groups 1 and 2 and a 5-percentage point growth in groups 1 and 2 in 2011 and 2012 when compared to 2010. In 2013, the growth was in group 4 and 5 indicating that more households were moving longer distances.

Figure 3.11 features the spatial distribution of the distance decay coefficient for 1990, 1995, 2000, 2005, 2010, and 2015. Households in western portions of the country, counties

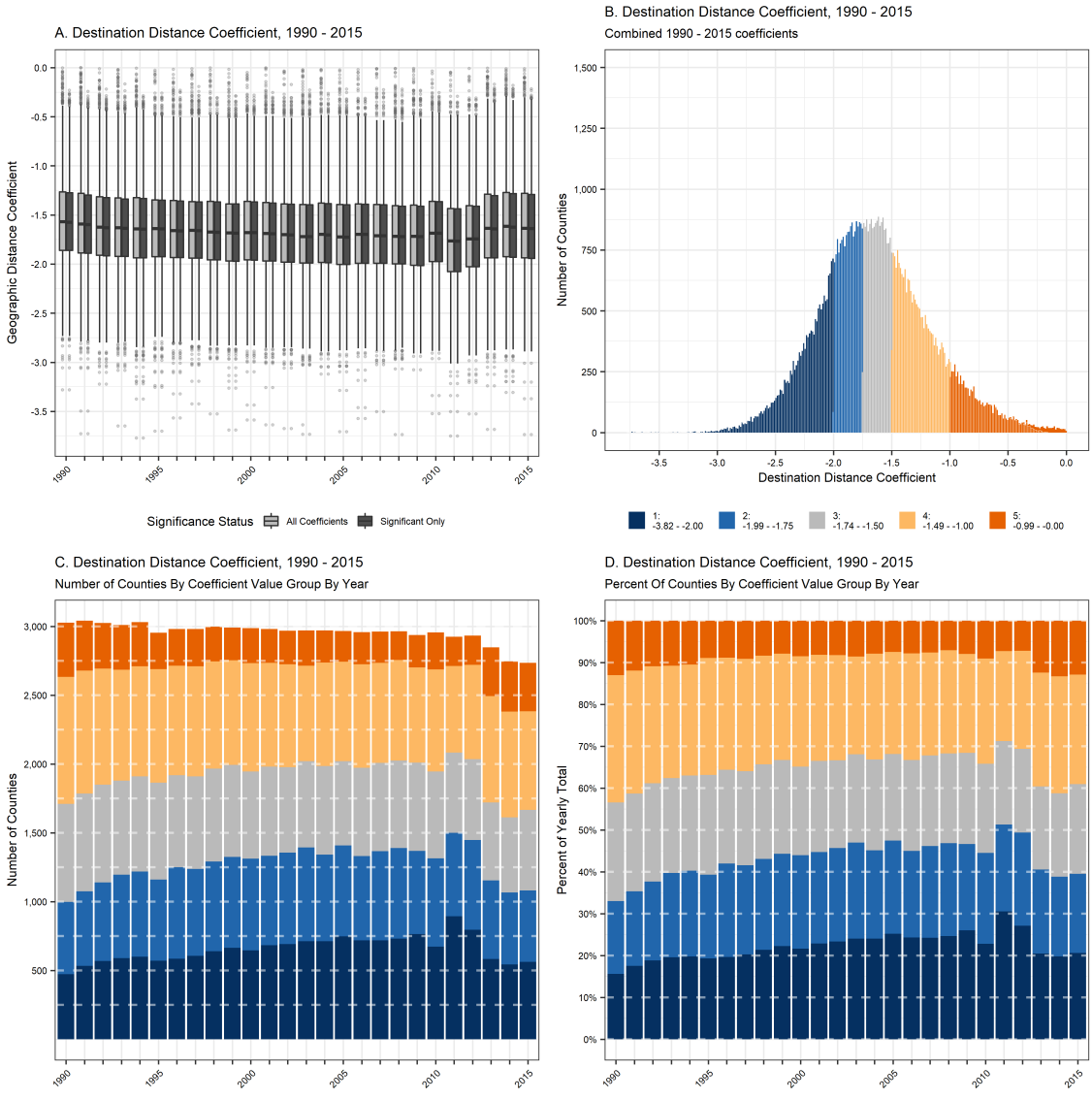


Figure 3.10: Statistical distributions of the destination distance coefficient, 1990-2015

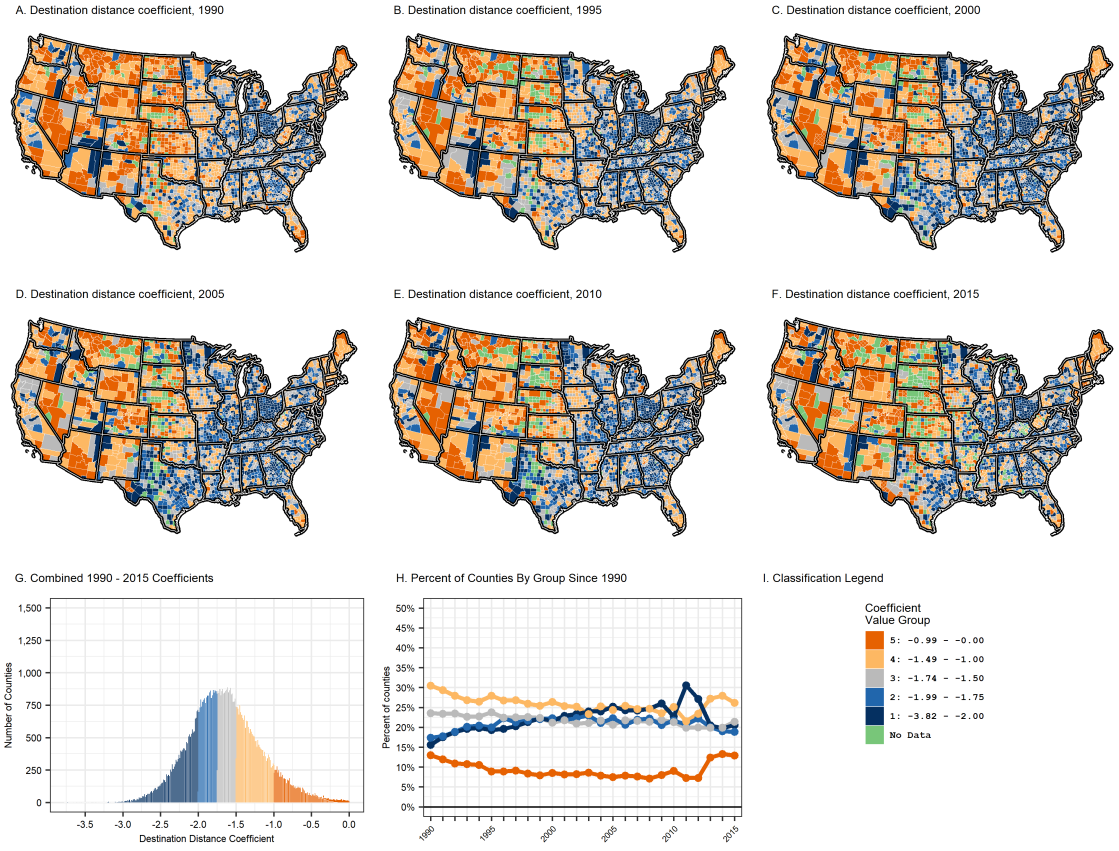


Figure 3.11: Spatial distributions of the destination distance coefficient, 1990-2015

in Florida, and counties in Maine prefer moving greater distances. This is in some effect due to the nature of these counties: they are simply further from other parts of the US and by that fact the households that leave counties in Florida and Maine must move further. However, households in some rural counties in western states such as Arizona, California, Oregon, and Washington show preference for shorter distance. Households in counties in the rust belt, Texas, and the south showcase a preference for shorter distance moves. There are more opportunities for shorter distance moves as counties in these states are smaller and there is less distance between the metropolitan areas. Graphic H in Figure 3.11 expresses the same data as in graphic D in Figure 3.10 but in a line plot. Graphic H shows that more households are moving shorter distances from 1990 through 2011. The trend reverses in 2011 and more households start moving longer distances again.

3.4.4 Accessibility

Figure 3.12 features graphics pertaining to the destination accessibility coefficient. The interpretation of the destination accessibility measure is that positive values indicate households move to counties that are near other counties with higher population counts. Negative values indicate that households are moving to counties that are more distant from other population centers. Graphic A shows that the boxplots of the pooled all-coefficient distribution is different than the distribution of the significant coefficients only (at the 0.05 level) boxplot. Of the 77K accessibility coefficients, 87-percent are significant. Most of the non-significant values are in groups 4 and 5, around zero. The accessibility coefficient ranges from -5.58 to 2.27 and over time it appears that most households prefer less accessible destinations. The mechanics behind this trend are better explained with the graphics in Figure 3.13.

Figure 3.13 on page 110 features maps of the spatial distribution of the accessibility coefficient for years 1990, 1995, 2000, 2005, 2010, and 2015. At first glance the figure is a bit counter intuitive. What is shown is that the accessibility coefficient is lowest in the rustbelt and eastern seaboard and generally greater in western counties. This is a trend consistent across time. The maps showcasing the distribution of the accessibility measure, Figure 3.5

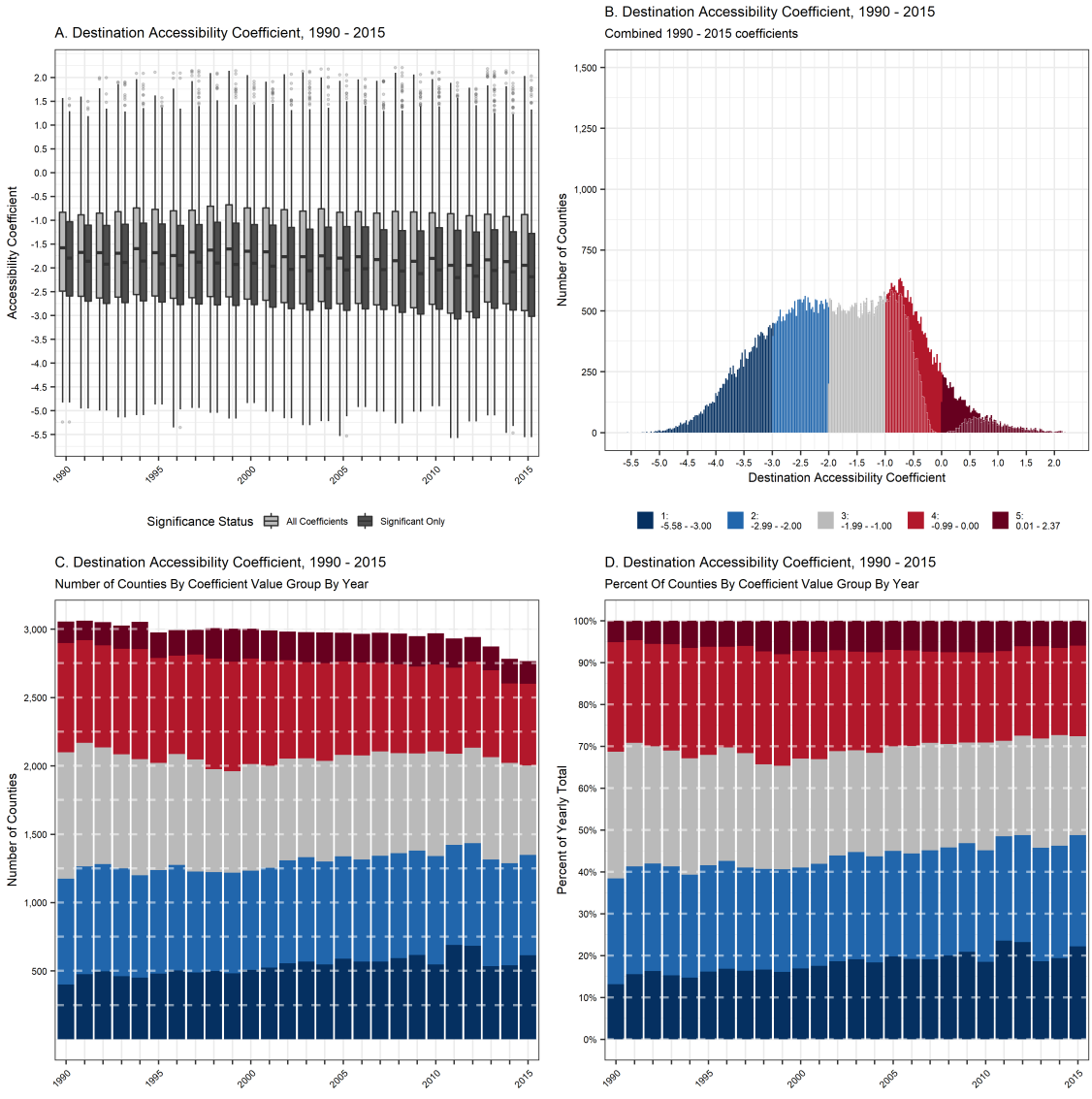


Figure 3.12: Statistical distributions of the destination accessibility coefficient, 1990-2015

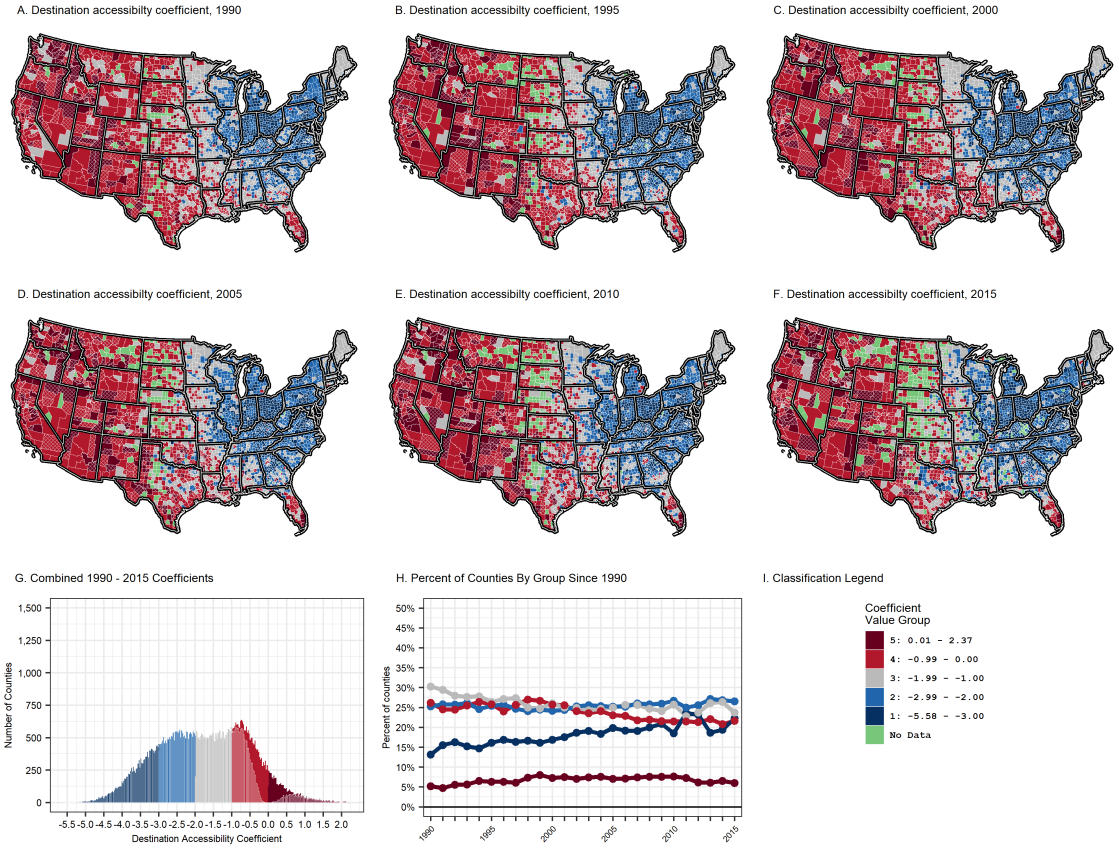


Figure 3.13: Spatial distributions of the destination accessibility coefficient, 1990-2015

on page 94, show that the accessibility measure (not the coefficient) is highest in the rust belt and eastern seaboard and lower in western portion of the US. The maps in Figure 3.13 are showcasing the relative nature of the spatial interaction models. One explanation is that a move that originates in a highly accessible area is more likely to terminate in a less accessible area and vice versa. The other explanation is that counties in California, Arizona, Texas, and Florida are destinations and these destinations have smaller accessibility measures. During the study's period, California is always a top ten destination from other states, Arizona is always one of the top 25 destinations from other states, Texas is a top 10 destination, and Florida is a top 10 destination for households from other states. Of note is that the majority of coefficients around zero, and the majority of the non-significant coefficients, are in the western portion of the US which indicates that for migration out of western portions of the country, accessibility is less of a factor considered in the selection of a destination. Another way to describe this relationship is that counties in the western portion of the country are generally larger in area and less dense while the reverse is true of counties in the eastern portion of the country. This means that counties in the West are less accessible than those in the East. As households in the West are generally moving around the West, they are moving to a relatively more accessible county. As households in the East are moving to the West, they are moving to a less accessible county.

3.4.5 Within-state migration

The within-state migration coefficient measures the relative preference that households have for in- versus out-of- state migration. Positive values indicate that an in-state destination is more preferential than out-of-state destination. In every year of the 1990 through 2015 period, the data in aggregate show that slightly more than 50-percent of flows are to a destination within the same state of origin. Graphic A in Figure 3.14 features two boxplots for each year. Values for the within-state migration coefficient range from -3.45 to 7.99 with the median value less than 2.0. The light grey boxplot shows all coefficients for all counties and the dark grey boxplot shows just the significant only coefficients. The median

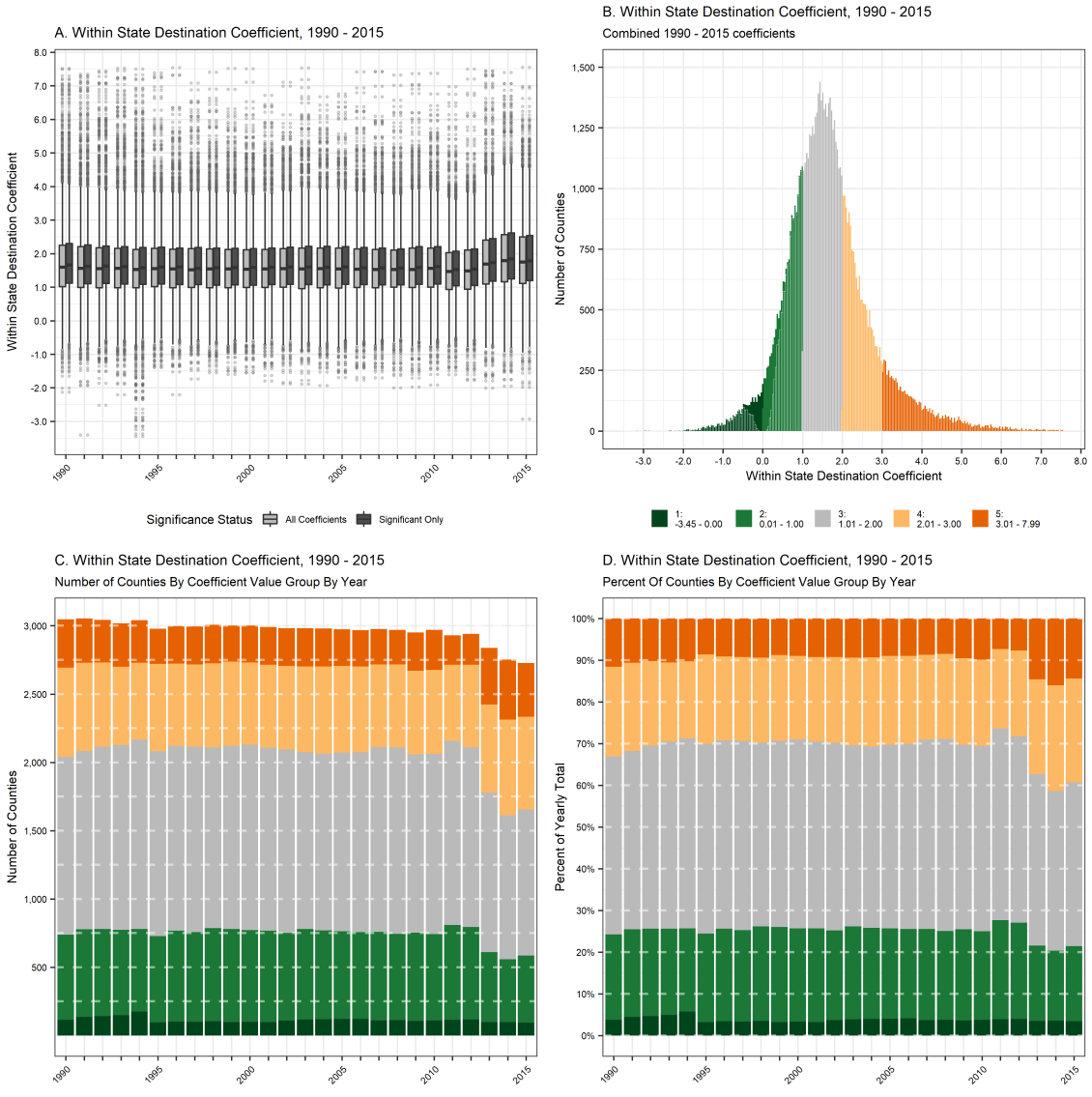


Figure 3.14: Statistical distributions of the within-state migration coefficient, 1990-2015

and interquartile ranges are very similar for years 1990 through 2010. In 2011 and 2012 the interquartile range shrinks slightly. Starting in 2013, the median value increases slightly, and the interquartile range expands indicating that the preference for within-state moves is more varied.

Graphic B features an all-years combined histogram of these data and shows that the non-significant coefficients are centered around zero. Negative values, group 1, dark green, indicate that households prefer out-of-state destinations. The light green color, group 2, indicates that within-state destinations are less preferred. Groups 3 through 5 are indicative of a preference for in-state destinations. Graphic C shows the counts of counties in each group by year showing that the trends are stable and graphic D expresses that information as a percentage. This is done to account for the change in the number of counties with outgoing flows. For groups three, four and five, an in-state destination is more desirable. In aggregate across the US, it looks like most moves are within-state. While many households prefer a within-state destination, the spatial interaction models showcase many counties with households expressing a preference for within-state movement, but not all of them. The bulk of coefficients are greater than 1.0 indicating that a within-state destination, for most households, is more desirable than not. State borders shape mobility.

Figure 3.15 features maps of the spatial distribution of the within-state destination coefficient for years 1990, 1995, 2000, 2005, 2010, and 2015. Households in California, portions of eastern Texas, Florida, North Carolina, and Ohio desire out-of-state destinations, areas shaded in green. For households in counties in these states, an in-state destination is not as desirable as an out-of-state destination. Conversely, for households in counties in Montana, Idaho, Nevada, Utah, Colorado, and Maine, in-state destinations are more desirable.

3.5 Examining the destination preferences of migrating households

In this chapter I have shown how to fit origin specific production constrained spatial interaction models using county-to-county household migration data for each year in the 1990 through 2015 period. Four covariates were used in the construction of these models to illus-

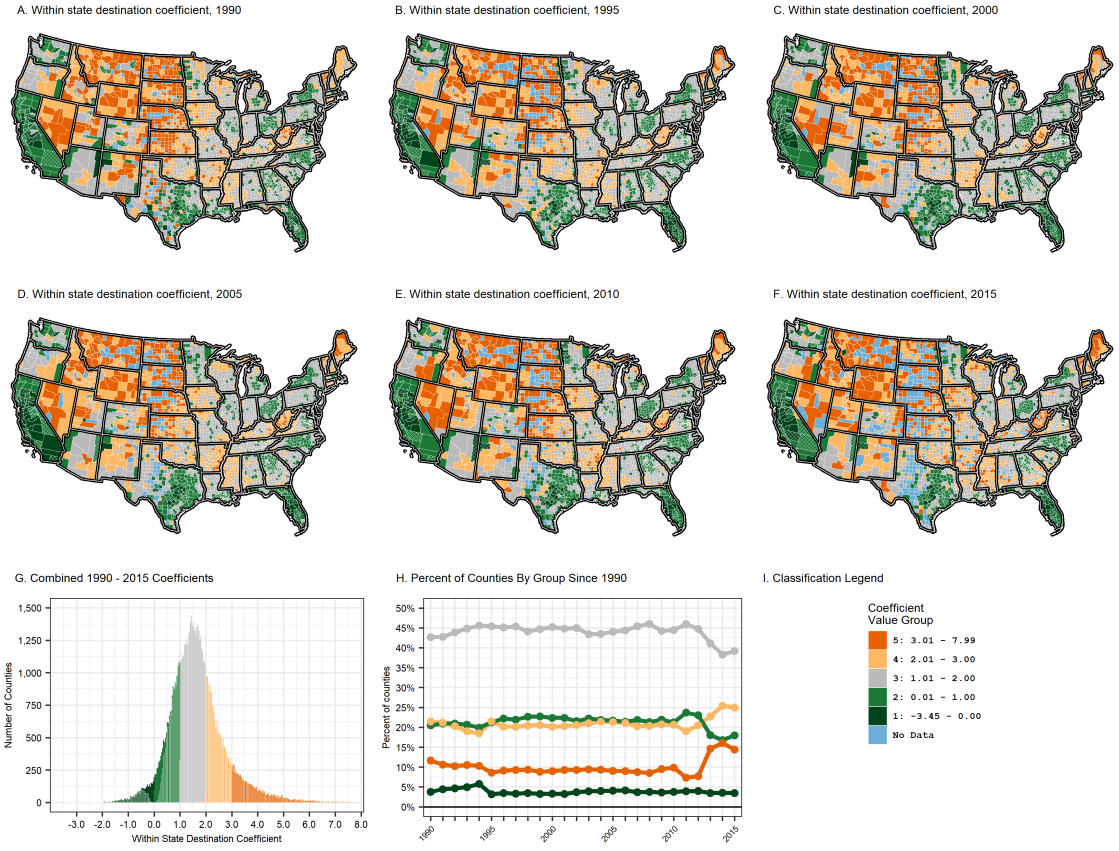


Figure 3.15: Spatial distributions of the within-state destination coefficient, 1990-2015

trate the destination preferences of out-migrating households for each county: the number of non-migrant households in the destination, the destination's distance from the origin, the destination's accessibility, and whether the destination is within the state of origin. A series of graphics were employed to showcase how these four covariates illustrated the preferences of outgoing households. A consistent five-category grouping and color scheme was applied to the all-years combined distribution of each coefficient enabling a comparison across time. Graphics illustrating the statistical and spatial distributions of the coefficients with the same color scheme and grouping were utilized showcasing the dynamics of each internal migration component.

The growth in the lowest groups of the destination mass, accessibility, and distance coefficients is in part due to a result of the growing US population and constant distance between US counties. The These trends are a result of the prominence of destinations in California, Arizona, Texas, and Florida. While the US population has increased since 1990, the population growth has not been uniform. In addition, there are effectively the same number of counties since 1990 which means that the number of origin and destination pairs are fixed which means that for any given county, there are only a fixed number of destinations an outward-bound migrant can select.

The growth in the lower group of the destination mass coefficient is indicative of larger counties becoming larger and smaller counties not growing as fast or at all. Because the same relative numbers of households are moving and are therefore going to the same destinations and the same number of destinations, there is growth in the lower groups of the distributions. If households were only going to metropolitan areas, more values greater than 1.0 would be visible. If households were migrating to more rural areas, more counties would feature coefficients less than 1.0.

The accessibility measure has more to do with the stability of county borders, which counties saw increases in population, and the spatial configuration of those counties. Again, the coefficients are measures of relativity and not an absolute measure. That is, coefficients less than 1.0 prefer less accessible areas and coefficients greater than 1.0 prefer more accessible

areas. This is why the maps in Figure 3.5, page 94, and in Figure 3.13, page 110, feature inverted color gradients. Blue counties are indicative of households moving to less accessible areas such as households in eastern portions of the county moving west, which generally feature less accessible areas.

The distance and the within-state migration indicator variables warrant joint discussion. With respect to the distance component of these models, the trend over time is one of households exhibiting increasing preference for shorter distance moves indicating that more households are moving shorter distances. The within-state migration indicator showcases the spatiality of migration trends. For households in some counties, a within-state destination is simply not as desirable as an out-of-state destination. Over time, more households in California, Texas, and Florida, in particular, are selecting out-of-state destinations. Households leaving California are preferring further out-of-state destinations while households leaving Texas and Florida are preferring closer out-of-state destinations. Conversely, in Idaho, Montana, Wyoming, and Nevada, households are showing a preference for moves within the same state of origin. This might seem at odds with the distance decay coefficient, but when considering the relative areal size of the counties and therefore the areal size of the states, the understanding is that households in these states are moving longer distances within the state of origin.

I have shown how four covariates can be used to illustrate internal migration trends and how using the same four covariates across time can illustrate the dynamics of internal migration. Of course, as good as this chapter is, it does have its limitations. This chapter is a study of migration in aggregate. While the IRS county-to-county migration data feature a fantastic amount of spatial and temporal detail, other than the gross aggregate income variable, there are no details pertaining to the socioeconomic and demographic characteristics of migratory households. Accordingly, it is unknown how much of this is migration across the life course. For example, students migrating to colleges and universities, people migrating due to the location of specific industries (military bases and resource extractive industries, for example). The next chapter of this dissertation addresses internal household migration

within the context of the life-course via age structures and other place-based characteristics of the origin and destination.

Chapter 4

DETERMINANTS OF METROPOLITAN AND NON-METROPOLITAN HOUSEHOLD MIGRATION, 1990-2015

Since the 1970s, there have been multiple periods where non-metropolitan counties grew more quickly than metropolitan counties (Fuguitt and Beale 1996), suggesting evolving spatial forms of internal migration. This chapter investigates the spatiality and the scale of internal migration by examining the determinants of four distinct types of household movement during the 1990 through 2015 period: non-metropolitan-to-non-metropolitan movement, non-metropolitan-to-metropolitan movement, metropolitan-to-non-metropolitan movement, and metropolitan-to-metropolitan movement. Socioeconomic, demographic, and outdoor amenities are used to explain these four different types of household movement and in turn augment the discussion on internal migration in a period of decline (Cooke 2013; Molloy et al. 2011).

A gravitational theory of human movement posits that migration between any two areas is a function of the origin's mass, the destination's mass, and the distance between the origin and the destination areas. This macro-behavioral understanding of human movement treats migration as a function of origin and destination characteristics: local area characteristics motivate, or push, migrants out of the origin and destination characteristics attract, or pull, migrants to the destination. The next section of this chapter describes periods of non-metropolitan migration and growth in juxtaposition with population concentration in metropolitan areas. This is done to highlight and accentuate the different types of household movement between counties and look at migration beyond declining migration rats. Next, this chapter describes the data used to investigate the determinants of four different types of household movement: the enhanced county-to-county household migration data and several

publicly available datasets with commensurate spatial and temporal resolution. The third section of this chapter describes the modelling technique used to ascertain the determinants of migration and the fourth section provides visualization and discussion of the results.

4.1 Metropolitan and non-metropolitan migration: a brief history

In summarizing his laws of migration, Ravenstein (1889, p. 287) noted that towns increase in population at the expense of the rural parts of the country. Towns grow in population through migration more quickly than by natural increase. A little more than eighty years later, Zelinsky's (1971) mobility transition framework conceptualized human mobility at different stages of societal development. The framework featured five stages with each stage of mobility a reflection of increased development. The first stage in the mobility transition described limited mobility while the second and third stages described massive movement from rural areas to cities. The fourth stage notes that most migration is interurban and intraurban and the fifth stage details that nearly all migration would be urban-to-urban and the majority of the population would be in urban areas[1]. Population concentration was the norm. Migration trends in the 1970s through the early 1980s and late 1980s through mid-1990s, however, showed that non-metropolitan areas grew faster than metropolitan areas indicating that migration was (and is) more than population concentrating in dense urban areas. Indeed, dense cities in the 1990s declined in population (Glaeser and Shapiro 2003).

Ballard and Fuguitt (1985) identify four periods of growth and decline from 1900 through 1980. The period from 1940 through 1960 was marked by suburbanization and rural decline and the period from 1960 through 1980 was marked by population deconcentration - people moved out of dense, urban areas. Researchers in the 1970s and 1980s documented the population deconcentration phenomenon and coined several terms for it: a "clean break" from the previous regime of metropolitan expansion (Vining and Strauss 1977), the "non-metropolitan turnaround" (Fuguitt 1985), and the "migration turnaround" (Agresta 1985).

[1]See Cooke et al. (2018) for a review of the historical impact and relevance of Zelinsky's (1971) migration transition.

The term "counterurbanization" is also used in conjunction with population deconcentration (Berry 1976; Mitchell 2004) as is "rural gentrification" (Phillips 2009).

Broadly, these concepts refer to population migrating from metropolitan areas to non-metropolitan areas. In general, researchers agree that the population deconcentration trend first noticed in the 1970s ended in the early 1980s. Using data available at the time, Agresta (1985) and Long and DeAre (1988) concluded that the turnaround was temporary and that population was again concentrating in urban areas (Frey 1993). Evidence from the mid-1980s through the mid-1990s, however, suggested that the US underwent another round of nonmetropolitan growth.

Fuguitt and Beale (1996) examined data from 1970 through 1994 and find two distinct periods of non-metropolitan growth, 1970 through 1980 and 1986 through 1994. The period 1981 through 1985 saw declines in non-metropolitan population. The period of non-metro growth from 1986 through 1994 was attributed to favorable economic conditions enabling people to select less-populated places (Long and Nucci 1998) and the rise of information technologies enabling remote work (Beyers and Lindahl 1996). While non-metropolitan counties experienced net in-migration in the mid-1990s, smaller metropolitan areas also experienced growth in population.

Plane et al. (2005) examined migration between metropolitan areas during the 1995 through 2000 period. The authors classify metropolitan and non-metropolitan areas into one of seven, ranked, categories indicating a hierarchy of urban form based on population size. Movement between the ranks was examined using a migration effectiveness calculation, giving an indication of the degree of interchange between the population size ranks. The authors found that the most unbalanced flows were down the urban hierarchy, indicating a degree of population deconcentration in the mid-1990s. Manson and Richard E. Groop (2000) found similar results in movement down the urban hierarchy in the mid-1990s and that migration from large central cities to adjacent suburbs led to increases in economic disparity.

Using IRS county-to-county migration data for the 1990 through 2010 period, DeWaard

et al. (2020a) examine the internal migration rate via four different movement types: non-metropolitan-to-non-metropolitan, non-metropolitan-to-metropolitan, metropolitan-to-non-metropolitan, and metropolitan-to-metropolitan. The author's find that changes in the internal migration rates of the four movement types are a function of the spatial interconnectivity of the migration system (the number of county-to-county ties). Over time, the metropolitan-to-non-metropolitan migration rate decreased and the metropolitan-to-metropolitan migration rate increased. The decreasing metropolitan-to-non-metropolitan migration rate during the 1994 through 2010 period and the increasing non-metropolitan-to-metropolitan migration rate in the 1995 through 2007 period suggest that population is concentrating.

Cycles of metropolitan and non-metropolitan growth are visible in the IRS's county-to-county household movement. Figure 4.1 features the net migration rates for metropolitan and non-metropolitan areas for each year during the 1990 through 2015 period. There are two sets of migration rates. Graphic A features migration rates holding the 1990 metropolitan categories constant and graphic B features migration rates using the most current metropolitan definitions for each year. Throughout this chapter and where prudent, I will showcase rates and counts using the 1990 metropolitan categories and the contemporary year definitions. This is to underscore how information about population and household movement is dependent upon the metropolitan classification scheme. However, all subsequent analysis is done using the fixed 1990 metropolitan categories. The 4.2.1 section, page 129, describes in further detail metropolitan and non-metropolitan classifications.

Graphic A in 4.1 shows that the non-metropolitan migration rate was increasing from 1990 through 1994, decreasing from 1995 through 2000, increasing again from 2001 through 2005, decreasing from 2006, 2007, 2008, a brief increase in 2009, and then decreasing and generally flat beginning in 2010 through 2014. In 2015, the non-metropolitan migration rate was again increasing. Graphic B, using the current year metropolitan definitions, shows a similar shaped curve, but with one key difference: the positive net migration rate seen in 2001 through 2006 using the 1990 metropolitan definitions is removed when using the contemporary year definitions. This indicates that non-metropolitan growth in the 2000s

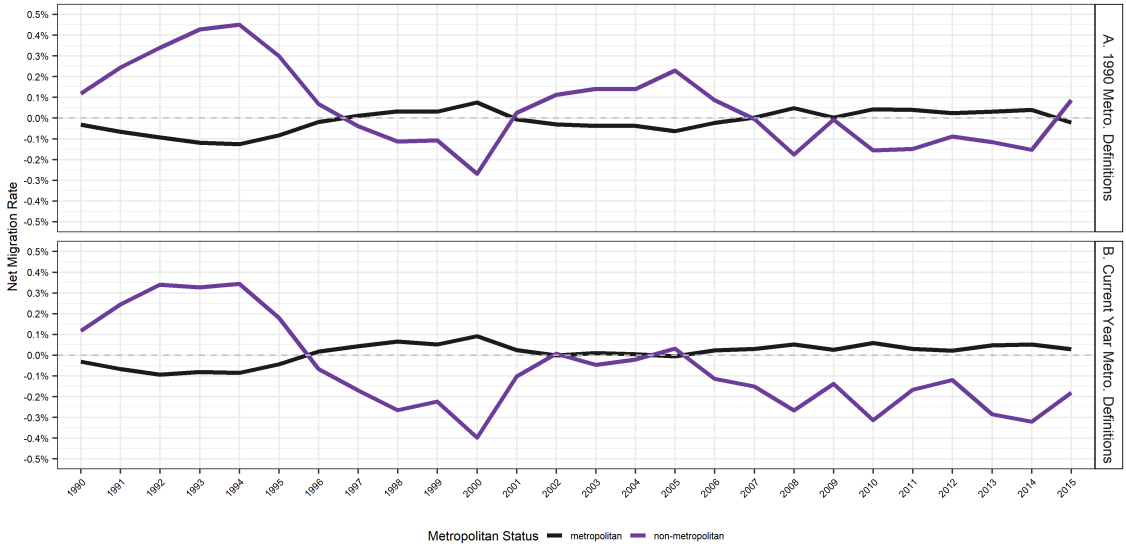


Figure 4.1: Net migration rate by metropolitan status, 1990-2015

was to areas that were reclassified as metropolitan in 2000.

Some rural areas are experiencing depopulation due to chronic net out-migration coupled with natural decrease (Johnson and Lichter 2019). Young people of child-bearing age migrate away from non-metropolitan areas leaving behind an older population approaching end-of-life. Johnson and Lichter highlight how some rural counties have experienced population increases on account of Hispanic in-migration, amenity, and retirement migration. Retiring baby-boomers migrating to non-metropolitan counties generate demand in local services which in turn attracts Hispanic in-migration (Nelson et al. 2009). McGranahan et al. (2011) illustrate the conditions driving outmigration from non-metropolitan counties and find that low population density, isolation, and lack of outdoor amenities promote outmigration. The authors also state that rural areas in general lose young adults and gain families and retirees. A host of destination amenities are examined during retirement moves by Duncombe et al. (2001) and Taylor (2011) and both studies find that migration is driven in part by a desire for a location in an area rich with amenities yet close enough to maintain contact with networks developed in more urban areas. Examining the life-history of migrants returning

to rural communities, Reichert et al. (2014) and Wall and Von Reichert (2013) find that the reasons for return to rural areas are family, career, and community bonds even though lack of rural employment opportunities presents barriers for some people to return. Other researchers have investigated return migration to rural areas in Sweden (Lundholm 2012) and Northern Ireland (Stockdale et al. 2013) and found that attachment to place (as examined through place of birth in the case of Sweden) and childhood memories (in the case of Northern Ireland) influence migration to rural areas. Implicit in these studies is that migration to and from non-metropolitan areas is a function of the age of the migrant.

The regularity of the age profile of migration is well documented (Castro and Rogers 1984; Little and Rogers 2007; Pandit 1997; Plane 1993; Plane and Heins 2003; Plane and Rogerson 1991). Migration rates peak in early adulthood and a second, smaller peak is seen in retirement age. Johnson et al. (2005) examine spatial and temporal changes in age-specific net migration rates over a period beginning in the 1950s and ending in the 1990s. Not only do age-specific migration rates vary by decade, but migration rates also vary by the metropolitan status of the county. Migration rates peak for people in their late 20s in core metropolitan counties, across multiple periods. In non-core metropolitan counties, new metropolitan counties, and non-metropolitan counties, migration rates peak for people in their early 30s (Johnson et al. 2005). Migration at specific ages is not simply a function of the migrant's age, it is frequently a function of the migrant's stage in the life course stage. Indeed, it has long been shown that the propensity to migrate is a function of the migrant's stage in the life course (Rogers et al. 2002) and mobility coincides with major life course transitions (Clark and Withers 2009). Bernard et al. (2014) examine life course transitions and the age profile of internal migrants in a little over two dozen countries. In the cross-national comparison of internal migration rates, the authors show that migration age profiles mimic the age structure of key life-course transitions such as completing education, entering the labor force, partnering, and having a child. The authors do state that life-course transitions do not always include migration and vice-versa. Part of Bernard et al.'s 2014 work echoes Roseman's (1983) and Miller's (1973) finding that employment reasons are

the biggest driver of interstate migration, at least during the 1970s.

The causes of the periods of growth and decline in the net migration rates shown in Figure 4.1 and in previous research is a subject of discussion as reasons for migration to metropolitan and non-metropolitan areas vary. Williams and Sofranko (1979) find that quality of life is a factor determining nonmetropolitan in-migration. Dillman (1979), in a review of the literature, found that Americans would prefer to live in more rural areas, but commuting and employment constraints necessitated a residential location in more urban areas. Long and DeAre (1988) find that nonmetropolitan growth is sensitive to business cycles. Others cite amenities, such as opportunities for outdoor recreation and pleasant weather, as reasons for migration, often in retirement, to non-metropolitan areas (McGranahan et al. 2011; Rappaport 2007). Graves (1979) looks at each destination city as supplying bundle of goods - labor market vitality, income, and climate amenities - and examines migration between metropolitan areas as tradeoffs of those goods. Movement is made to maximize purchase of those goods, with amenities a strong determinant of migration. Greenwood and Hunt (1989) critique Graves' conclusion and find that employment matters more. Though in a follow up study, Greenwood et al. (1991) find that amenities are in part driving migration to western and southern states. Scott (2010) finds that employment opportunities matter more than amenities for working age engineers.

Several of the determinants for a decline in migration are also determinants of movement. In explaining declining migration rates, Kaplan and Schulhofer-Wohl (2017) find that the decline in migration is due to a decline in geographically specific returns to wages and Rupasingha et al. (2015) find that the determinants of metropolitan-to-non-metropolitan and non-metropolitan-to-metropolitan migration include population density, industry mix, natural amenities, and the percentage of older people explain the trends in migration in both the 1995-2000 period and the 2005-2009 period. Withers and Clark (2006) find that changes in housing affordability after a move influence the participation of women in the labor force. Using data pertaining to the 1986 through 1993 period, the authors find that for families moving to more affordable areas, wives are more likely to exit the labor force and conversely,

enter the labor force when moving to a less affordable area. Sasser (2010) also showcases how housing affordability is an increasing factor in the determinants of internal migration for the years 1977 through 2006 using IRS data. Metropolitan and non-metropolitan net migration rates go through periods of decline and growth and the destinations that migrants select is a function of the life-course. Concluding their paper with a remark about the new forms of migration in the US, Plane et al. (2005, p. 15318) state that "[m]igration reflects qualitatively and quantitatively different determinants at the various key stages of the life course. No longer are most migrants excess laborers in rural areas flocking, by steps, into rapidly industrializing cities; today's U.S. internal migrants are people who find themselves in the wrong places at the wrong times in their lives."

The resolution of this spatial and temporal mismatch, migration, is evident in Figure 4.1, page 122. This chapter contributes to the discussion on the spatially of internal migration by looking beyond migration rates and instead at the determinants of movement between, to, and from metropolitan and non-metropolitan counties as functions of the age structure and other place-based characteristics of the origin and destination counties. Several studies have focused on describing the rates of metropolitan (Plane et al. 2005) and non-metropolitan migration (Fuguitt and Beale 1996). And several of the previous studies have examined the spatial and temporal variation in age specific component of migration (Johnson et al. 2005). And several studies have identified the determinants of migration (age, economic indicators, and amenities. Finally, DeWaard et al. (2020a) use an identical movement classification scheme to investigate the decline in migration as a function of spatial interconnectivity. This research investigates an aspect of the spatiality of internal migration by examining the determinants of four different types of internal migration over time.

The geography of the outcome of the four types of movement, net migration, is visible in the maps in Figure 4.2, page 126. The seven maps (graphics A through G), bar plot (graphic H), and accompanying legend (graphic I), illustrate the spatial and temporal trends in net migration which is the sum of households entering a county minus the sum of households exiting a county. Net migration is disaggregated by household flows to and from

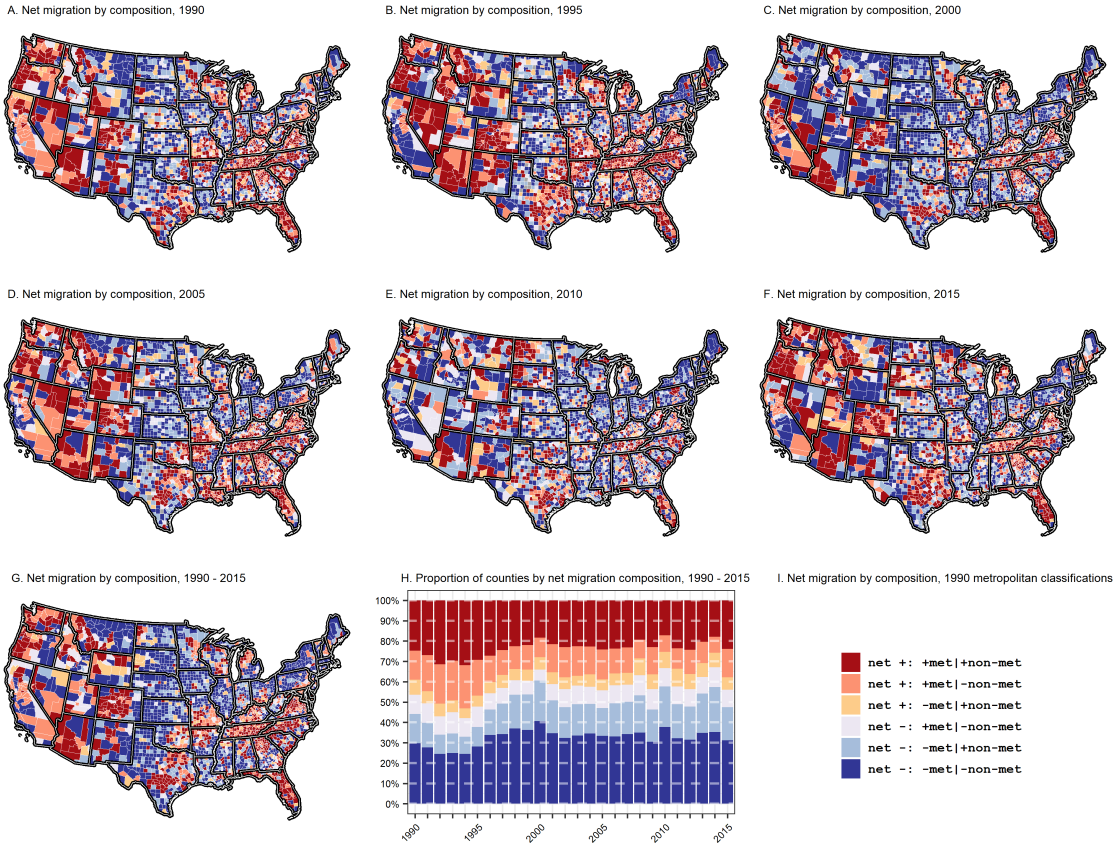


Figure 4.2: Net migration by composition, 1990-2015

non-metropolitan counties and household flows to and from metropolitan counties. Dark red indicates that a county had a positive net migration rate and drew a surplus of households from both metropolitan and non-metropolitan counties. Lighter shades of red still indicate a positive net migration rate, but a component of the net migration composition is negative. For example, a county could have a positive metropolitan net migration rate and a negative non-metropolitan migration rate, but more households came from metropolitan counties than from non-metropolitan counties. A county shaded dark blue indicates that more households moved to both non-metropolitan and metropolitan counties than households moved in from non-metropolitan and metropolitan counties. Lighter shades of blue indicate an overall negative migration rate, but some component is positive. There are six categories total: three positive and three negative.

Graphics A through F show the net migration by composition for years 1990, 1995, 2000, 2005, 2010, and 2015. Graphic G is the 26-year total of net-migration by composition. Graphic I features the legend and graphic H features a bar plot with the proportion of the counties in each category over time. With the exception in the early 1990s, more counties featured negative net migration rates than positive net migration rates indicating that population is concentrating. More counties featured positive net migration in the early 1990s indicating that population was deconcentrating. This period coincides with the positive non-metropolitan net migration rate seen in Figure 4.1, page 122. The greater the share of negative net migration rates, the greater the degree to which population is concentrating. Most population gain is from households migrating from metropolitan counties and most population loss is from households migrating to metropolitan counties. This is in effect, migration up and down the urban hierarchy (Plane et al. 2005). There are small numbers of counties where households increased in number from non-metropolitan origins and decreased in number to metropolitan origins. The distribution of this concentration is west-ward and to the southerly and it is driven by households moving out of counties in the rust belt in upstate New York, Pennsylvania, Ohio, Indiana, and Illinois and migration out of counties in rural states such as North Dakota, South Dakota, Nebraska, and Iowa.

The gains and losses featured in Figure 4.2, page 126, represent households moving between metropolitan and non-metropolitan counties. This chapter investigates the determinants of that movement. Using the enhanced county-to-county migration data described in chapter two, I estimate the yearly determinants of migration across four types of movement: non-metropolitan-to-non-metropolitan movement, non-metropolitan-to-metropolitan movement, metropolitan-to-non-metropolitan movement, and metropolitan-to-metropolitan movement. A limitation of the IRS county-to-county household migration dataset is that it is not possible to unpack the four types of mobility by demographic and socioeconomic characteristics. For example, the age and sex or the educational attainment of the members of the household are not known in the IRS county-to-county household migration. Certainly, much could be gained from an analysis incorporating those characteristics. For example, Wright and Ellis (2019) show that migrants with a STEM degree (science, engineering, technology, and mathematics) undergraduate degree move further and are less influenced by destination amenities.

The enhanced county-to-county household migration data do offer tremendous temporal and spatial resolution, however. The opportunity provided by the county-to-county household migration data - yearly measurement of household movement between consistent spatial units over a 26-year period - present a simultaneous challenge in sourcing data with commensurate spatial and temporal resolutions. While previous research has identified several factors explaining internal migration, those variables are often not available for each year in the 1990 through 2015 period. Through the curation of publicly available data matching the temporal and spatial resolution of the enhanced county-to-county migration data and the application of a spatial interaction modelling framework, I will show how four types of household movement are functions of origin and destination place-based characteristics. To that end, this chapter will now turn to a discussion on the data used to model the four types of county-to-county household movement.

4.2 Data

Several publicly available data sets power the analysis of this chapter. The numbers of households moving between counties come from the enhanced IRS county-to-county household movement data. Of the numerous determinants of internal migration identified by previous research, the following determinants that can be sourced from publicly available data with spatial and temporal resolution matching the enhanced county-to-county migration data are age structure (US Census Bureau), economic indicators (Bureau of Labor Statistics), and outdoor amenity indicators (USDA and NOAA). In addition to these determinants, several geographic variables are included to account for the spatial configuration of origin and destination counties. Finally, the metropolitan and non-metropolitan classifications of the origin and destination counties are used to identify the four different types of movement. Identifying metropolitan and non-metropolitan counties is a critical and necessary step in this analysis and a discussion of the metropolitan classifications is where I will begin the discussion on the data in use in this chapter.

4.2.1 Metropolitan and non-metropolitan classifications

The Office of Management and Budget (OMB) (2010) currently defines a metropolitan statistical area (MSA) as a densely settled area with at least 50,000 people. The MSA comprises the central county or central counties plus adjacent, outlying counties with a high degree of social and economic integration as measured through commuting behavior. The basic concept of the MSA has existed since the 1950s and the standards for defining metropolitan counties have changed which in turn has changed the manifest of metropolitan counties (Adams et al. 1999; Farley 2007; Klove 1952). In 2003, the OMB introduced the concept of a micropolitan county (Office of Management and Budget 2000) which is a county featuring an urban core with a population between 10,000 and 50,000 people. Ratcliffe (2002) notes that the OMB views the metropolitan and micropolitan concepts as having urban form, urban function, and a high degree of socio-economic linkages between the urban core and the outer

edges[2]. Metropolitan and micropolitan counties feature urban aspects but are not expressly and entirely urban. In fact, the Office of Management and Budget (2000, p. 82228) explicitly states that "The Metropolitan and Micropolitan Statistical Area Standards do not equate to an urban-rural classification; all counties included in Metropolitan and Micropolitan Statistical Areas and many other counties contain both urban and rural territory and populations." A non-metropolitan county is a county that is not metropolitan because it lacks the features of a metropolitan county. A non-metropolitan county is defined by what it lacks.

As the standards for defining a metropolitan county have changed over time, so too has the manifest of metropolitan counties. Accordingly, a county can gain or lose metropolitan or micropolitan status. During the 1990 through 2015 period, metropolitan designations changed 12 times (US Census Bureau 2016, 2013, 2010, 2009, 2007, 2006, 2005, 2004, 2003, 1999, 1993, 1990). During this period, 40-percent of all counties maintained a consistent non-metropolitan designation, 23-percent of counties maintained a consistent metropolitan designation, 17-percent transitioned form non-metropolitan to micropolitan, and 8-percent transitioned from non-metropolitan to metropolitan. The remaining 12-percent of counties fluctuated between a metropolitan, non-metropolitan, and micropolitan designation. Grouping the micropolitan category with the non-metropolitan category removes the inconsistency. After applying this grouping, 60-percent of counties featured a non-metropolitan designation and 23-percent featured a metropolitan designation. Eight-percent of counties transitioned from non-metropolitan to metropolitan. The remaining 9-percent of counties fluctuated between metropolitan and non-metropolitan. Whether combining the micropolitan and non-metropolitan classification or not, in both cases, a small percentage of counties switched from metropolitan to non-metropolitan. Using contemporary metropolitan classification versus a constant metropolitan classification has ramifications for the counts of people in each type of area as well as the trends in the growth in non-metropolitan areas.

---

[2]See Morrill et al. (1999) for an alternative urban and rural classification scheme and Cromartie and Bucholtz (2008) for a discussion of how rural areas are defined. Isserman (2005) discusses the implications of urban and rural designations for research and public policy.

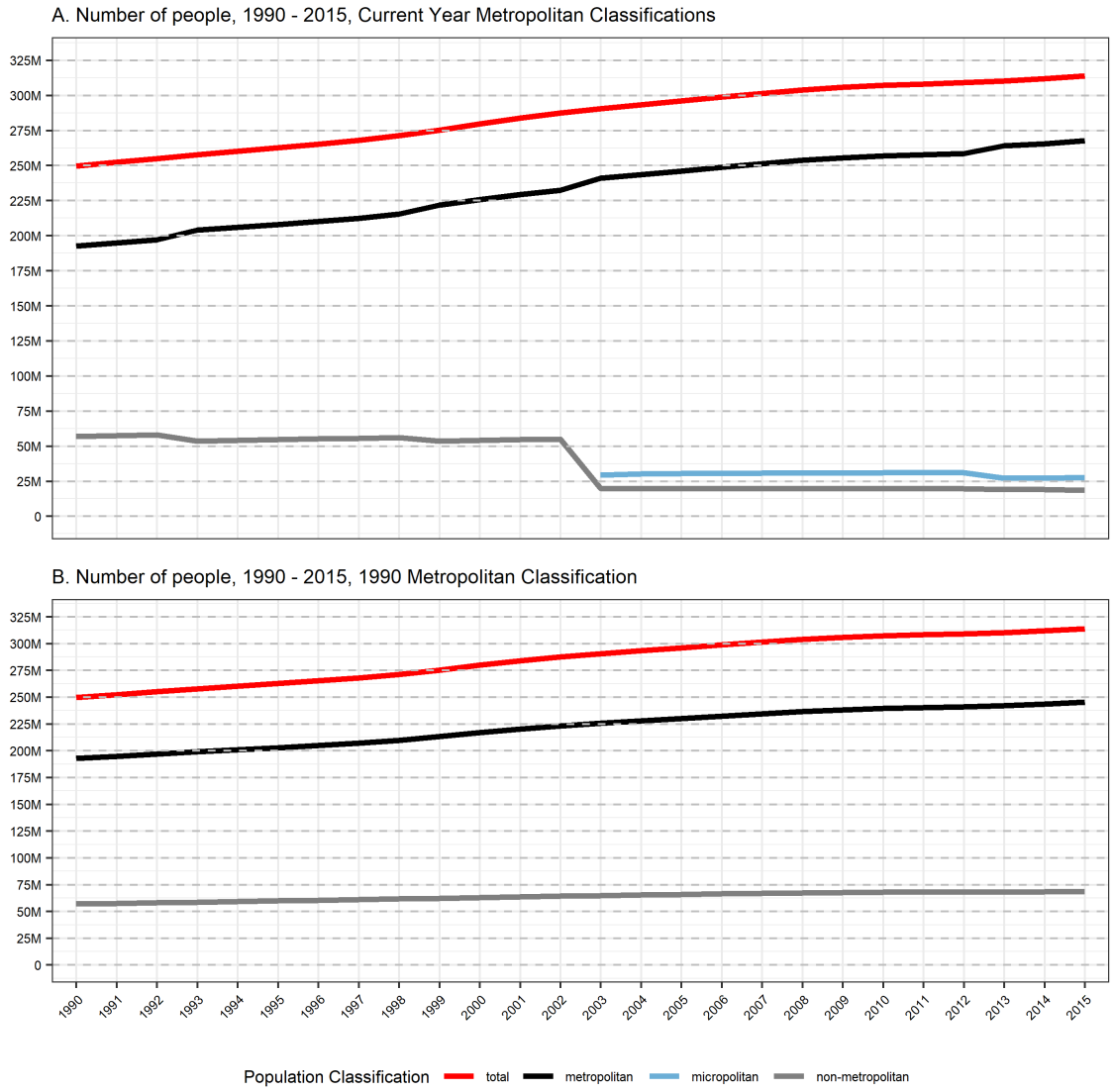


Figure 4.3: Number of people by different metropolitan classification schemas, 1990-2015

Figure 4.3 illustrates the effect of using each year's then-current metropolitan classification, graphic A, and the 1990 metropolitan classification applied to each year, graphic B. The same data are used in each graphic, only the metropolitan classifications change. In 1990, 77-percent of the population in the US - 193 million people - were in a metropolitan area. Using the current-year classifications, that number increased to 85-percent, 267 million people. In the 1990 to 2015 period, using contemporary metropolitan, micropolitan, and non-metropolitan definitions, the number of people in non-metropolitan counties decreased from 56.7 million people to 18.5 million people. Fixing the metropolitan county definitions to 1990 (and in the process combining the non-metropolitan and micropolitan categories), the non-metropolitan population increased to 68.3 million people. Examining the growth of the metropolitan population in the US since 1960, Johnson and Lichter (2020) find that the increase in the metropolitan population was not due to growth in established metropolitan areas, it was in the establishment of new metropolitan areas. That is, a formerly non-metropolitan county becomes metropolitan due to reclassification by the OMB and in the process millions of people are added (instantly) to the metropolitan population register. A county receives a metropolitan designation by reaching a certain population threshold by in-migration, often from metropolitan counties. Before I can describe the trends present in both the county-to-county movement data and the numbers of people by age, it is necessary to first establish a consistent set of metropolitan definitions going forward. As Johnson and Lichter (2020) illustrate and as I have shown as well, the reclassification of non-metropolitan counties to metropolitan counties has been a significant driver of metropolitan growth. From 1960 to 2017, 753 non-metropolitan counties were reclassified as metropolitan, shifting nearly 70 million residents into a metropolitan status by 2017. The growth of the share of the US population that is considered metropolitan is due to reclassification[3].

[3]Metropolitan classifications are not static. A recent report from the Office of Management and Budget Office of Management and Budget (2021) recommend increasing the minimum population threshold for metropolitan counties to be increased from 50,000 to 100,000. This would shift 19 million people from residing in a metropolitan county to residing in a non-metropolitan county (Pipa and Geismar 2021). The 2021 proposal revisits an alternative delineation scheme from the OMB (1998) suggesting three categories based on population size. Metropolitan counties would feature more than 100K people, mesopolitan would

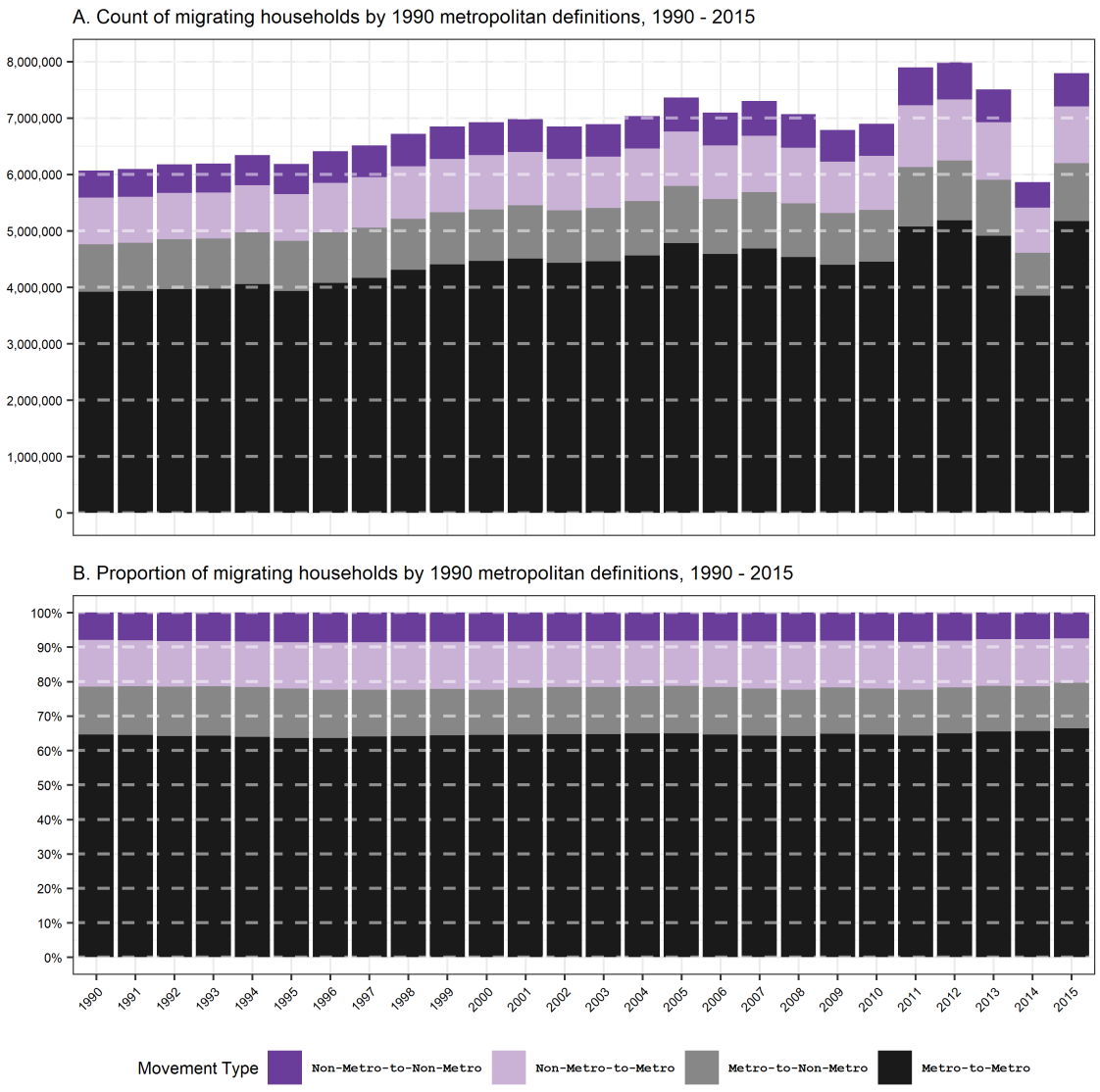


Figure 4.4: Migrating households by 1990 metropolitan definitions, 1990-2015

The dependent variable in all models is the number of households moving from origin county i to destination county j in year t; this variable comes from the enhanced IRS county-to-county migration data. The four different movement types are determined by assigning the 1990 metropolitan classification to each origin county and each destination county in each year in the 1990 through 2015 period. Figure 4.4 features the counts, graphic A, and proportions, graphic B, of migrating households by the 1990 metropolitan definitions for each year during the 1990 to 2015 period. The four movement types are non-metropolitan-to-non-metropolitan, non-metropolitan-to-metropolitan, metropolitan-to-non-metropolitan, and metropolitan-to-metropolitan. The number of households migrating generally increases over time. Small dips are seen in 1995, 2002, 2008, 2009, 2013, and 2014. Aside from 2013 and 2014, these dips are generally attributed to recessions. The dips in 2013 and 2014 are the result of differences in data processing. The share of each movement type remains remarkably consistent across the study period. On average, two-thirds of household movement is metropolitan-to-metropolitan. The proportions of non-metropolitan-to-metropolitan movement and metropolitan-to-non-metropolitan are each approximately 14-percent and the remaining 6-percent of moves are non-metropolitan-to-non-metropolitan. The enhanced county-to-county migration data enable an examination of migration to and from metropolitan and non-metropolitan counties, usually a difficult task due to limited non-metropolitan data sources (Foulkes and Newbold 2008).

A final visualization to underscore the differences between the four types of movement is the degree of movement between origin and destination counties when distance is measured by adjacency distance, or nodal distance, as opposed to a linear unit such as miles. An adjacency distance of 1 indicates that two counties are directly adjacent, such as Los Angeles County, CA and Orange County, CA. An adjacency distance of two indicates that the destination county is two counties from the origin county. For example, Los Angeles County, CA has an adjacency distance of two from San Diego County, CA, as Orange County is in between

---

feature a population count between 50K and 100K people and micropolitan would have a population count between 10K and 50K people.

Los Angeles County and San Diego County. Figure 4.5 on page 136 features the proportion of households by movement type by adjacency distance. The data in Figure 4.5 are the same as in Figure 4.4, just with an added distance component. Each of the four graphics in Figure 4.5 correspond to the different movement types. Adjacency values of one through five are singular values while values of 6 or more represent the grouped remainder. The average distance between any two directly adjacent counties in the contiguous 48 states is approximately 30 miles and, except for the grouped remainder category, each increment in adjacency distance is an increase of approximately 30 miles.

The proportions across the movement types are similar across time indicating that migration patterns are consistent. In each year across the four movement types, excluding the grouped remainder, the largest proportion of movement is between directly adjacent counties. There is a substantial difference in the proportion of households moving to a directly adjacent county and to a county that features an adjacency distance of two. Most moves are shorter distance across all movement types. The size of this share does differ by movement type, however. Graphic A features the proportion of households moving from one non-metropolitan county to another. In any given year, approximately 50-percent of households move between directly adjacent non-metropolitan counties. This trend does increase slightly over time. Graphics B and C, the proportion of households moving between non-metropolitan and metropolitan and metropolitan and non-metropolitan counties, respectively, showcase the most change. Over time, a larger share of households move between directly adjacent origins and destinations. Graphic D features the proportions of households moving between metropolitan counties. In any year in the 1990 through 2015 period, over 40-percent of moves are between directly adjacent metropolitan counties. This is similar to the distribution of movement between non-metropolitan counties: most moves are between directly adjacent counties. The four graphics in Figure 4.5 indicate that the immediacy of more local space satisfies the needs of mobile households. In addition, the proportion of households moving to directly adjacent counties has implications for the output of the regression models.

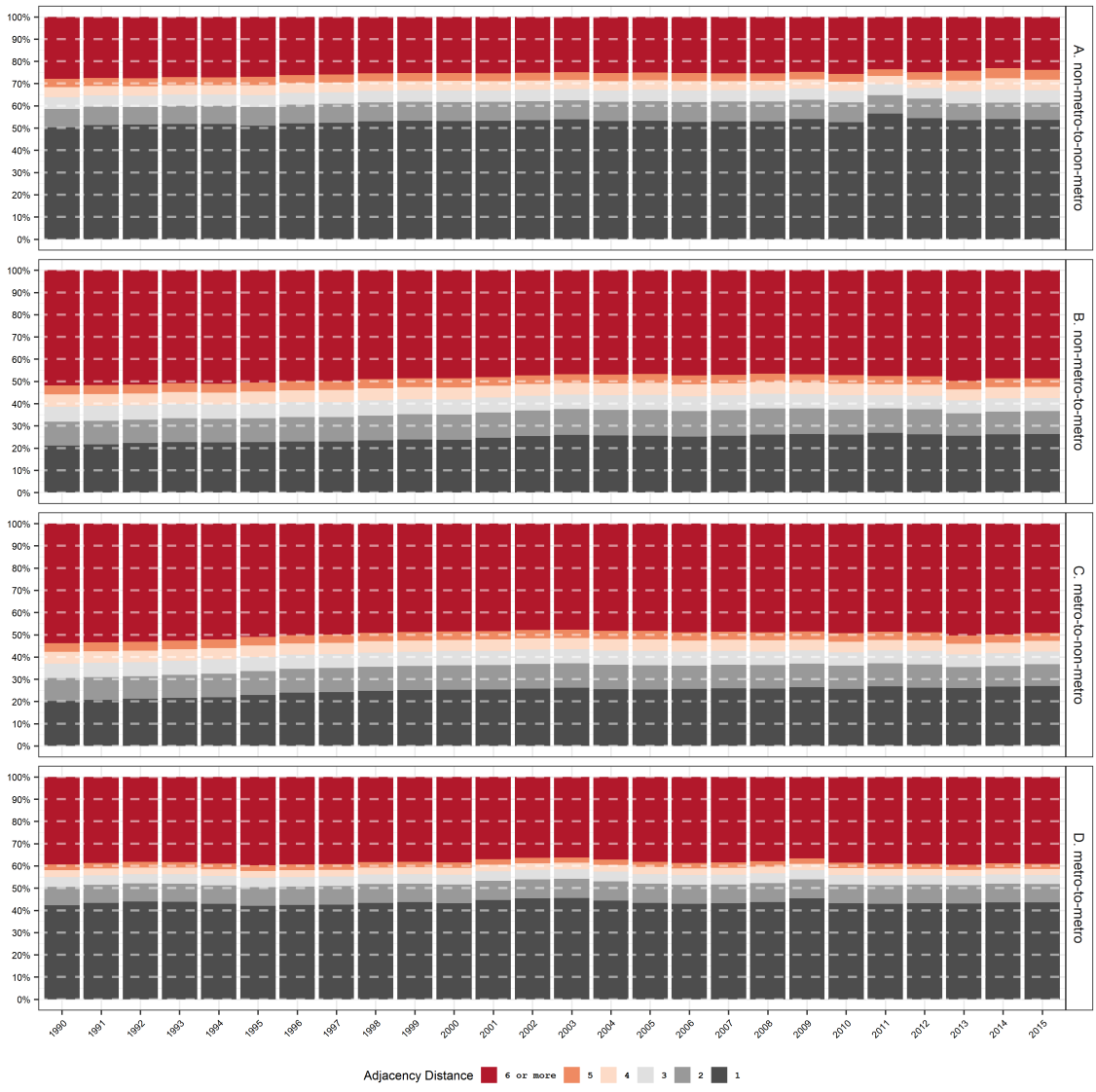


Figure 4.5: Proportion of moves by movement type by adjacency distance, 1990-2015, 1990 metropolitan definitions

4.2.2 Population age structure

The age data from the Census Bureau were specifically chosen because of its consistent spatial and temporal coverage: the same variables consistently tracking and measuring the same population component year after year at the same spatial resolution. While other county-level datasets are available, finding additional county level data both geographically complete (the data are available for all counties) and temporarily complete (the data are available for all years) is a challenge. The age structure variables come from the US Census Bureau's annual estimates of age by sex (Population Division 2020, 2016, 2000). These data, as initially downloaded, are presented in the standard 5-year age increments by sex. I aggregated the 5-year age groups by sex into the following combined-sex and age groups: ages 9 and younger, ages 10-19, ages 20-29, ages 30-39, ages 40-49, ages 50-59, and ages 60 to 70, ages 70-79, and ages 80 and older. The population in the US was approximately 250M in 1990 and increased by 64M to approximately 314M in 2015.

Using the 1990 metropolitan definitions, the number of people in metropolitan counties was approximately 193M and increased to 245M by 2015. The number of people residing in non-metropolitan counties, using the 1990 metropolitan definitions, was approximately 57M in 1990 and increased to 68M by 2015. Figure 4.6 on page 138 features three graphics showcasing the count of people by age group, graphic A, and the proportion of the population in each age group by metropolitan status, graphics B and C. Graphic A shows how the raw count people has increased over time. In general, metropolitan counties tend to skew slightly younger than non-metropolitan counties. Graphics B and C show the aging of the US. In 1990, the population under 40 in metropolitan counties was approximately 63-percent and 60-percent in non-metropolitan counties. By 2015, the population under 40 had dropped 10 percentage points for both metropolitan and non-metropolitan counties. Over 50-percent of the population is under 40 in metropolitan counties while in non-metropolitan counties, the population under 40 is 50-percent.

Each of the nine age groups are to be used as both origin and destination variables in the

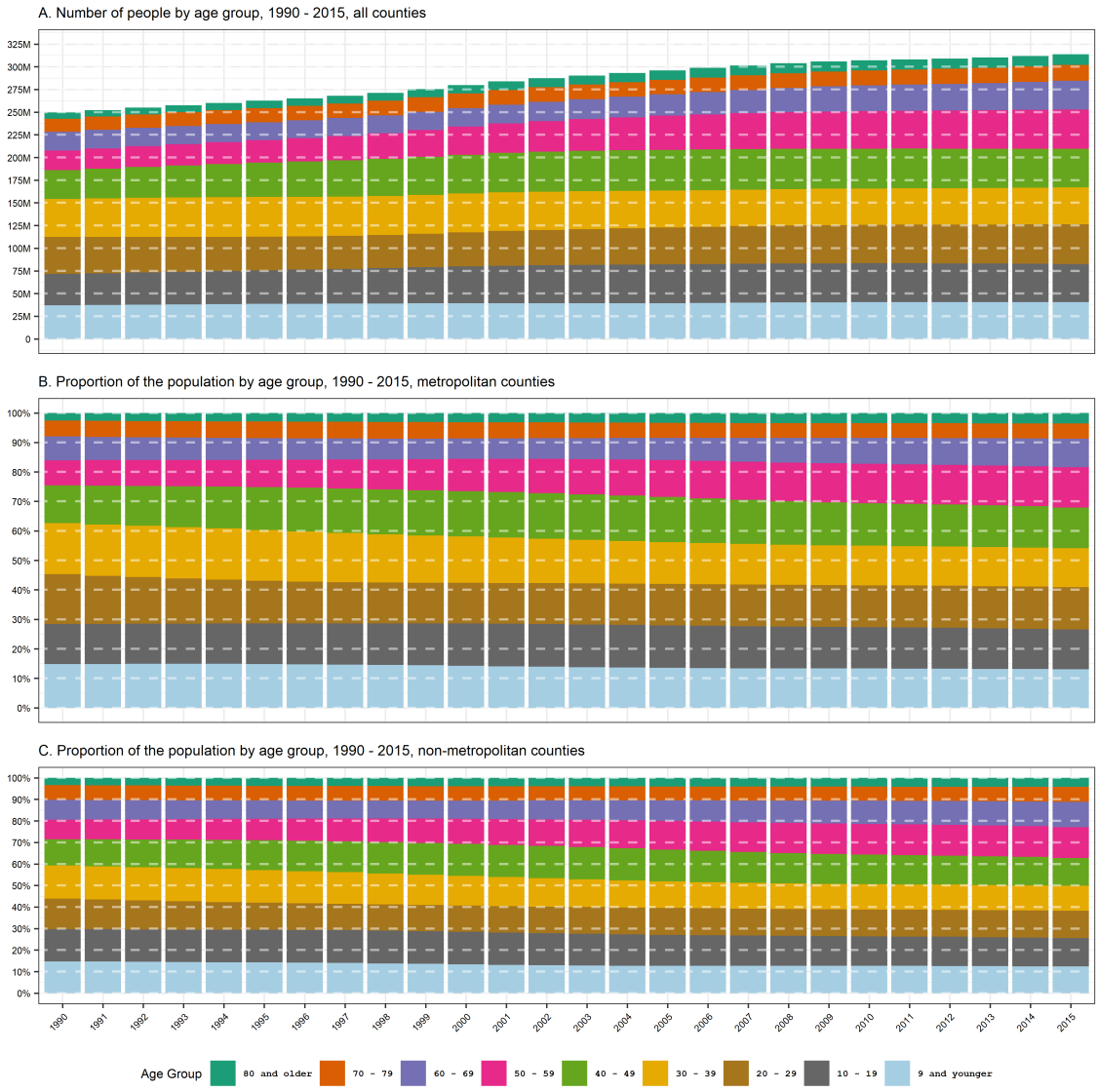


Figure 4.6: Counts and proportions of the population by age group and metropolitan status, 1990-2015

subsequent regression models. Structuring the regression models by these age groups enables the generation and comparison of age-specific population level push and pull factors. While a gravitational theory of human mobility suggests that increases in any one age group in the origin or the destination would increase push and pull factors respectively, previous studies have shown that mobility is differentiated by age group. Because of this, I assume that age differentials will show up in the regression coefficients.

I hypothesize the following rank order of each coefficient in any year for all four movement types: Ages 20-29 ¿ Ages 30-39 ¿ Ages 9 and younger ¿ Ages 60-69 ¿ Ages 40-49 ¿ Ages 50-59 ¿ Ages 10-19 ¿ Ages 70-79 ¿ Ages 80 and older as evidenced by Johnson et al. (2005). Given that most studies have shown greater rates of mobility in younger adults, I hypothesize that this group will feature the largest coefficients. The group with the second largest coefficient is the ages 30-39 group. I hypothesize that the ages 9 and younger group will feature the third largest coefficients on account of children of that age tend to live with younger adults. The ages 60-69 group will feature the fourth largest coefficients on account of the spike in mobility amongst retirement age adults. The ages 40-49 group and the 50-59 group will feature lower coefficients as this group is settling into careers. Similarly on account of having parents in the 40-49 group and the 50-59 group, the ages 10-19 group will feature lower coefficients. Finally, I expect the ages 70-79 and the 80 and older groups to feature the lowest coefficients. For movement originating in non-metropolitan areas, I hypothesize that younger populations will generate greater push factors. For movement terminating in non-metropolitan areas, I hypothesize that older populations will generate greater pull factors. The coefficients for older age groups for moves originating in metropolitan areas will generally be greater on account of older people moving out of metropolitan areas. For moves terminating in metropolitan areas, the coefficients for younger age groups will be greater.

4.2.3 Economic indicators

There are three economic indicators in use in this chapter. The first two, the county level unemployment rate and the county level annual average pay, both come from the Bureau

of Labor Statistics (BLS). The third variable, median house value, comes from the Decennial Census and the American Community Survey. The labor market data feature monthly estimates of the number of employed workers, unemployed workers, and the unemployment rate in each county in each year (Bureau of Labor Statistics 2020). The Local Area Unemployment Statistics program within the Bureau of Labor Statistics produces these estimates based on data from the Current Population Survey, the Quarterly Census of Employment and Wages, and the American Community Survey. Data from these sources, along with state unemployment insurance claims, are combined to produce the estimates of the number of unemployed workers. The multiple data sources in use enable the BLS to include agricultural workers, self-employed workers, unpaid family workers, and private household workers in the estimates. Because the unemployment data are at the county scale, I can visualize measures of central tendency for the distributions of metropolitan and non-metropolitan for each year in the 1990 through 2015 period. Figure 4.7 on page 141 features this information.

Graphic A in Figure 4.7 features the distribution of the number of unemployed people per county and graphic B features the distribution of the unemployment rate per county. Both graphics feature the distributions split by metropolitan counties, shaded dark grey, and non-metropolitan counties, shaded purple. The boxplots in each year show the median and inter-quartile range for each group of counties. A line plot was added to show the yearly average for each group. Metropolitan counties, on average, feature greater numbers of unemployed people compared to non-metropolitan counties. This is to be expected given the larger population size of metropolitan counties. The unemployment rate peaked in 1992, decreased throughout the 1990s, rose throughout the early 2000s, and decreased from 2003 through 2007. The unemployment rate peaked at approximately 7-percent in 2010 - the height of the great recession - and gradually declined. Some counties featured unemployment rates in the high teens in 2010. During the 1990 through 2008 period, the unemployment rate for metropolitan counties was less than the unemployment rate for non-metropolitan counties. Beginning in 2009 and lasting through 2012, the period corresponding to the Great Recession and the recovery, the average unemployment rate in both metropolitan and non-

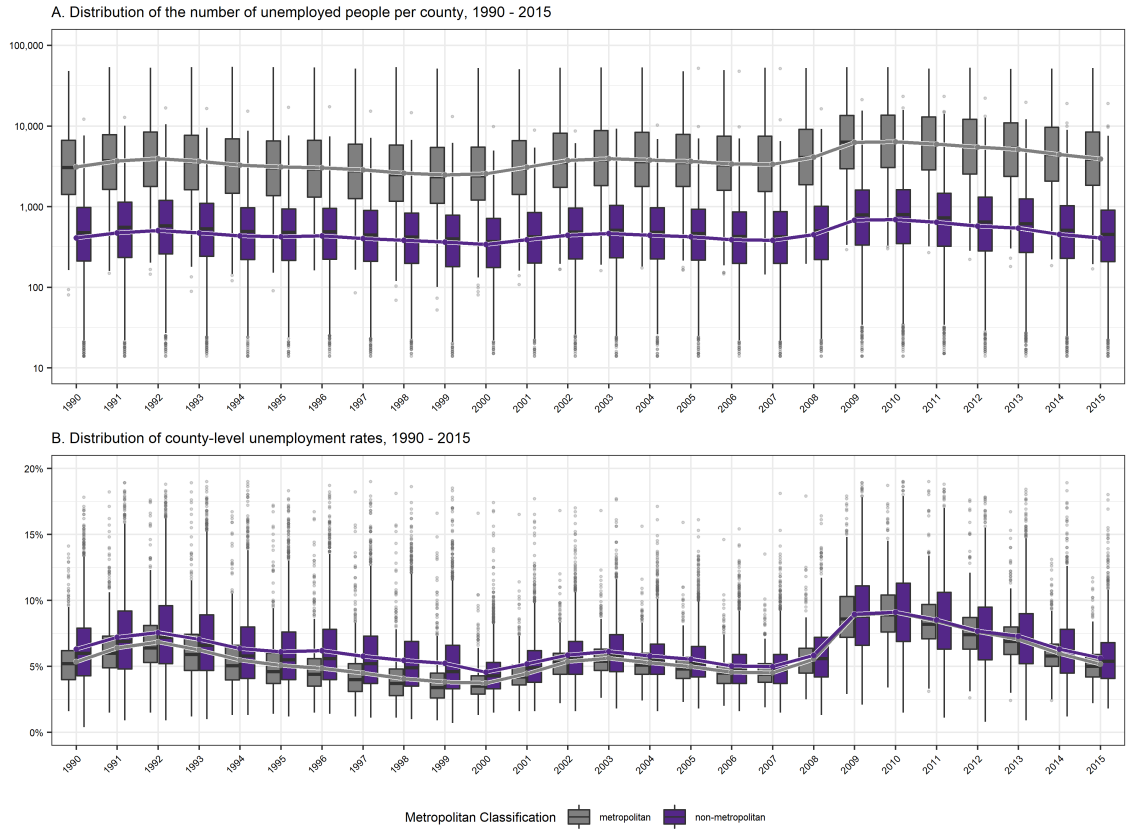


Figure 4.7: Distribution of the count and rate of unemployment, 1990-2015

metropolitan counties were approximately identical. By 2013, the unemployment rates were again diverging, and metropolitan counties featured lower rates of unemployment.

I hypothesize that high levels of unemployment will push people out of metropolitan and non-metropolitan counties and that lower levels of unemployment will pull people towards destinations. I hypothesize that the effect will be stronger in metropolitan areas. In each year in the study period, the distribution of the unemployment rate for metropolitan counties was more narrow than the distribution of the non-metropolitan unemployment rate. This research will show how differentials in unemployment rates translate to differentials in mobility determinants when examined across the four movement types. It could be that unemployment propels households in both metropolitan and non-metropolitan counties in a similar manner.

The Bureau of Labor Statistics' Quarterly Census of Employment and Wages (QCEW) provides quarterly counts of employment and wages as reported by employers (Bureau of Labor Statistics 2021). The primary source of QCEW data is administrative data from state unemployment insurance programs. Additional data sources include surveys conducted by the BLS. Together, the QCEW covers more than 95-percent of employment. QCEW data are available at multiple geographic scales and feature counts of employment and wages for about a dozen industries. Wages are reported in several forms: an annual aggregate total, an average annual weekly wage, and an annual average pay. I will use the annual average pay as an explanatory variable because this variable reflects the cost of labor as well as providing an indicator of the cost of living. Counties with higher annual average pay are assumed to have a higher cost of living. Similar to the BLS's measures of unemployment at the county level, I can showcase the differentials in annual average pay by metropolitan and non-metropolitan counties.

Figure 4.8 on page 141 features the distributions of the annual average pay in 2020 dollars for metropolitan and non-metropolitan counties. The average annual average pay for both metropolitan and non-metropolitan counties, represented in grey and purple, respectively, increased over time. The annual average pay is greater in metropolitan counties than

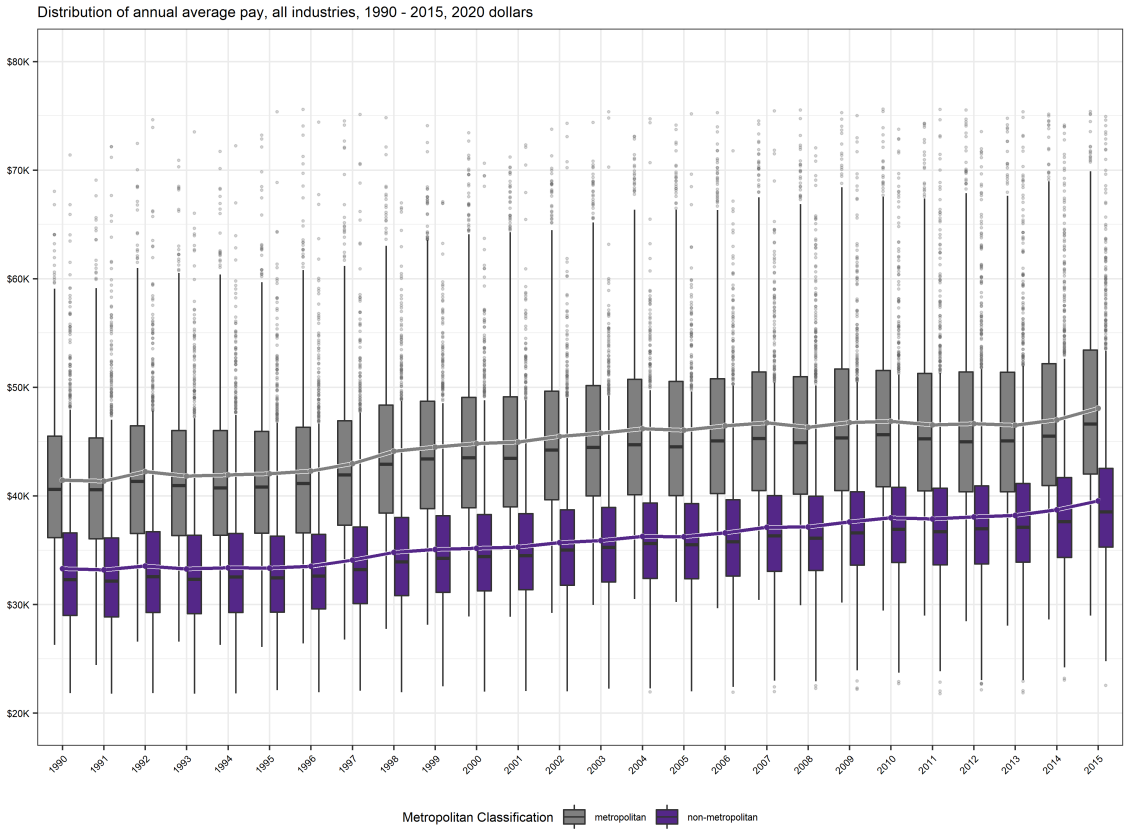


Figure 4.8: Distribution of annual average pay, 1990-2015

in non-metropolitan. From 1990 through 1994, the 25th annual average pay percentile in metropolitan counties generally matches the 75th annual average pay percentile. Starting in 1995 and ending in 2007, the 25th annual average pay percentile in metropolitan counties and the 75th annual average pay percentile in non-metropolitan counties were diverging, suggesting that wages in metropolitan counties were growing more quickly than in non-metropolitan counties. By 2008, the 25th and 75th percentiles in the metropolitan and non-metropolitan counties respectively were converging. In practice, this means that in any given year, the upper end of the annual average pay distribution in non-metropolitan counties is at parity with the lower end of the annual average pay distribution in metropolitan counties. Labor is paid more in metropolitan counties. Certainly, this is a function of the types of industries present in metropolitan counties and less a function of the metropolitan categories themselves. I hypothesize households will move from counties with lower relative annual average pay to counties with higher relative annual average pay. Households will be less likely to move to a county with a lower relative wage.

The final economic indicator, median house value is only available for years 1990, 2000, and a five-year average from 2008 through 2012. The 1990 and 2000 values come from the 1990 and 2000 Censuses of Population and Housing, respectively and the averaged 2008 through 2012 data come from the American Community Survey. All data were obtained through the National Historical Geographic Information System (Manson, Steven et al. 2020). Because the median house value of each county is not available for each year, I am carrying forward the values as supplied through decennial censuses for 1990 and 2000. Census 2000 was the last year of the 1-in-6 household long-form survey which generated detailed socio-economic data. Since 2009, data from the waves of the five-year American Community Survey have replaced the long form data. I am using the 2008 through 2012 wave because 2010 falls in the middle of the wave. The median home value for 2010 represents the average median home value during the five-year period beginning in 2008. Figure 4.9 on page 145 shows the distributions of median house value for 1990, 2000, and 2010 in 2020 dollars. The average median house value is greater in metropolitan counties.

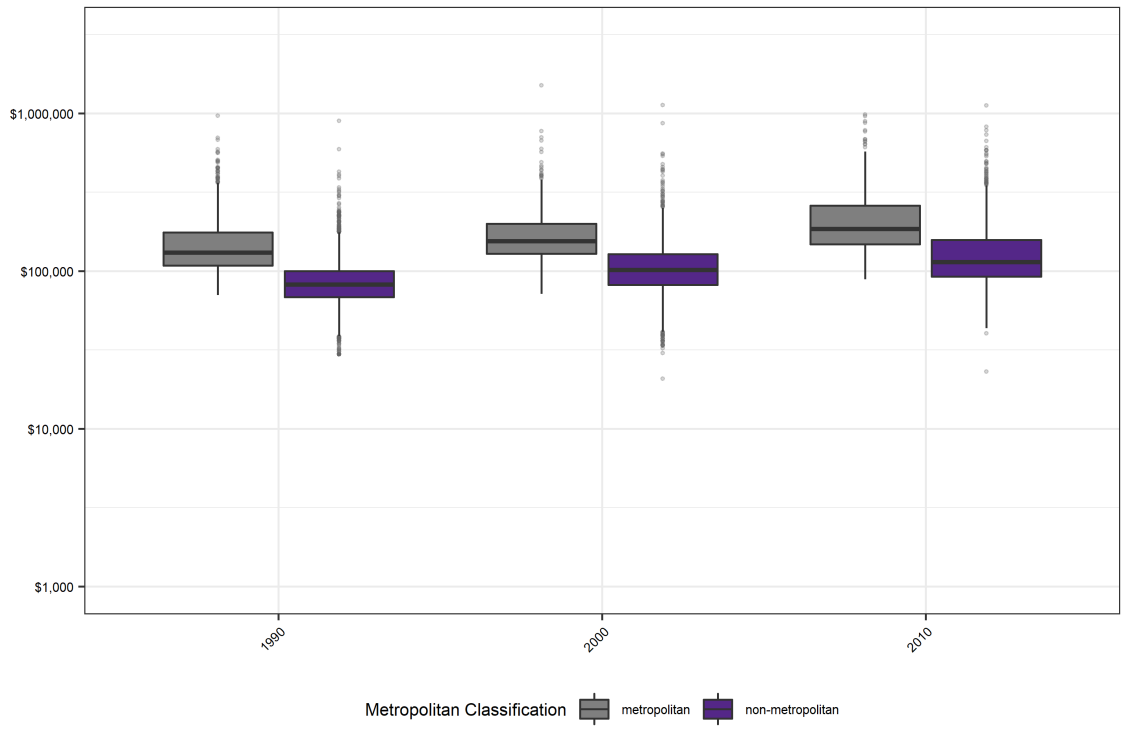


Figure 4.9: Distributions of median house value, 1990, 2000, and 2010, (2020 dollars)

4.2.4 Natural amenities

The variables used to investigate local area amenities are the number of inches of precipitation in Winter and the number of heating degree days (HDD) from the National Oceanic and Atmospheric Administration US Climate Divisional Database (Vose et al. 2014) and the USDA's natural amenity scale (McGranahan 1999). A heating degree day is the number of degrees below the average, usually 65-degrees Fahrenheit, in a one-day period (US Energy Information Administration 2021). For example, a hypothetical county has a temperature of 47 four days in a row. The difference between 65-degrees and 47-degrees is 18-degrees. Over the four-day period, the HDD for the hypothetical county is 72. I have aggregated the monthly counts to quarters to correspond to the seasons of the year. I will use measurements of HDD and total precipitation in January, February, and March to correspond to winter. These three months are chosen because these are bleak cold months in many areas of the US. In addition, results from the 1990s show that the weather drove population to warm, dry areas (Glaeser and Shapiro 2003). In addition, Wright and Ellis (2019), show that amenities do influence some migrants - those without a science, technology, engineering, or mathematics degree - and not others.

Figure 4.10 shows the distributions of the total number of inches of precipitation in Winter, graphic A, and heating degree days, graphic B, for metropolitan and non-metropolitan counties. While there are metropolitan counties and non-metropolitan counties throughout the country, suggesting a conclusion that weather in more southern counties would average out weather in more northern counties. This is not necessarily the case. On average, metropolitan counties feature a greater number of inches of precipitation than non-metropolitan counties and non-metropolitan counties feature a greater number of heating degree days than metropolitan counties. In general, in Winter, metropolitan counties are colder and wetter than non-metropolitan counties. This is reflecting the spatial distribution of metropolitan and non-metropolitan counties. There are many metropolitan counties in the Northeast of the US, famous for its cold and wet winters and many non-metropolitan

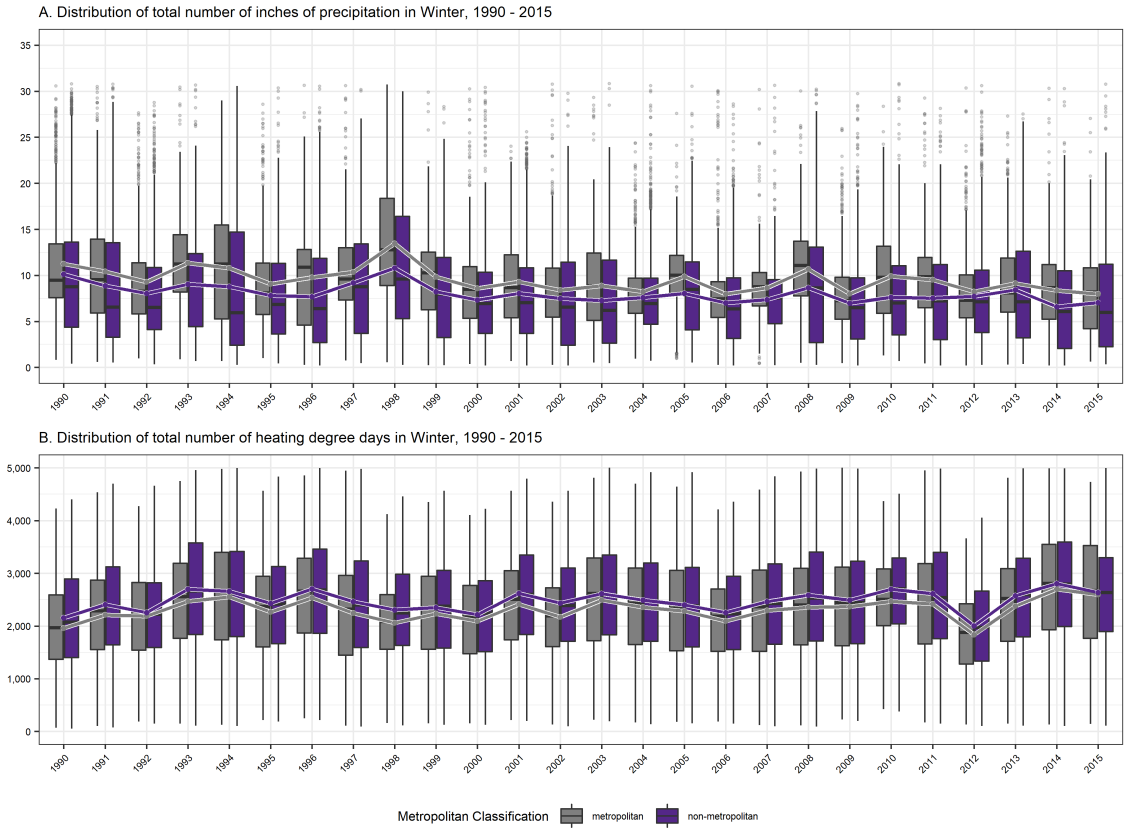


Figure 4.10: Distributions of the total number of inches of precipitation and heating degree days in Winter, 1990-2015

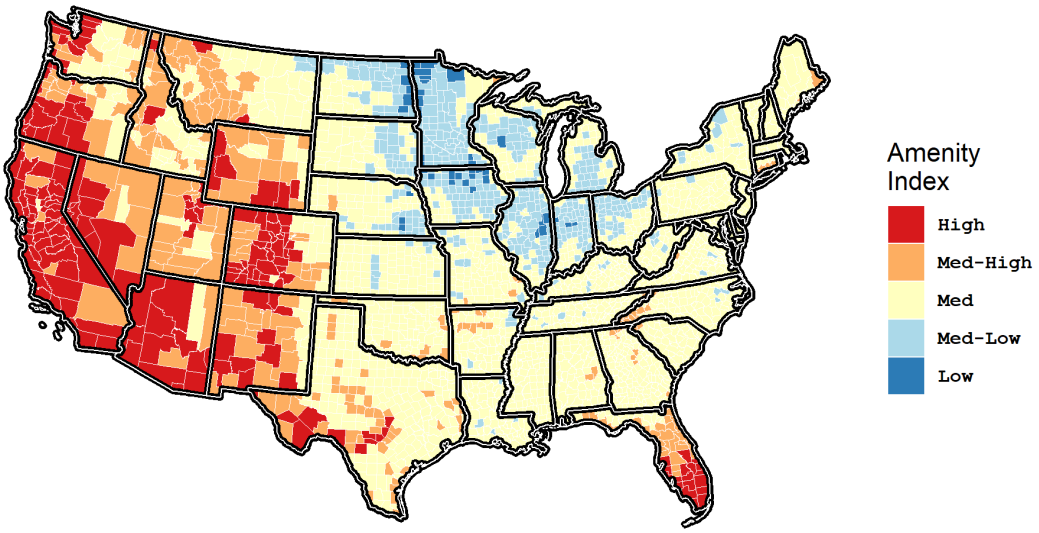


Figure 4.11: Spatial distribution of the amenity index

counties in arid and warmer portions of the western US. I hypothesize that increases in the number of heating degree days will increase outgoing migration from both metropolitan and non-metropolitan counties as will increases in the number of inches of precipitation. Households will move to warmer and drier areas.

The natural amenity index is produced by the Economic Research Service at the United States Department of Agriculture. The index is a composite of the types of outdoor and natural amenities people tend to find appealing (McGranahan 1999). The index includes average measures of temperature, the average hours of sunlight, the land surface form, and the percent of the county that is covered in water. The spatial distribution of the amenity index is displayed in Figure 4.11. Counties in the plains score lower on the amenity index than counties in the western portion of the US. It is assumed that counties with a low amenity index will increase flows of outgoing households and counties with a higher amenity index will attract households.

4.2.5 Distance and spatial structure

In addition to the age structure, amenity, and economic variables, I will include five geometric structure variables. Three of these variables are specific to the county-to-county pair and the other two are specific to each origin and destination county. The first geometric structure variable is the number of miles between the origin and the destination. This is the same distance measurement used in chapter three. Increases in distance mean less movement, *ceteris paribus*, and so this coefficient should always be negative. The second variable pertaining to geometric structure is a Boolean variable indicating whether not the origin and destination are directly adjacent. This variable is included to account for differences in county sizes and residential mobility. I expect this coefficient to always be positive indicating the desirability of a destination near the origin but offering something the origin does not currently feature. The third geometric structure variable is a Boolean variable indicating if the origin and destination are in the same state. I am including this variable for several reasons. First, the previous chapter demonstrated the utility of including this variable via an increased goodness-of-fit and an increase in the explanatory power of the models. Second, I am including this variable to show the importance of intrastate migration in relation to internal migration. State borders focus and direct migration and including this variable helps control for the effect of state borders. Over 50-percent of internal migration, as featured in the IRS county-to-county, is within-state (see Figure 2.5 on page 33 for a visualization). Third, I am including this variable to account for movement from one end of a state to another for particularly long moves. I hypothesize this variable to always be positive indicating that within-state destinations are more attractive than out-of-state destinations[4].

The final two geometric structure variables are the origin and destination accessibility measures, the same measure used in chapter three. This measure is conceptualized and operationalized in Yano et al. (2003) and is defined mathematically as:

---

[4]Distance is almost always included in models of this type. However, other researchers have found unique results when removing the nuisance factor of distance (Herting et al. 1997) or modeling distance as an outcome (Plane 1984).

$$A_j = \sum_{k(\neq j)} \frac{P_k}{d_{jk}} \tag{4.1}$$

The accessibility of area j is defined as the sum of the population of area k for all $k \neq j$ divided by the distance between j and k. This measure gives a sense of the total number of people-miles each area j is from other areas. Higher values of A_j indicate that j is more accessible to a population while lower values of A_j suggest that j is less accessible as d_{jk} grows larger. This variable is included because of its ability to control for the spatial structure of US counties, the distribution of population, and its ability to improve model fit. I hypothesize that for increases in origin accessibility and destination accessibility, the expected migratory outflow and the expected migratory inflow will decrease as these counties reflect nearness to other population centers and therefore greater access to amenities. Figure 3.5 on page 94 features the statistical and spatial distribution of this measure.

Concluding this section on the data is Table 4.1, which features the source and the expected sign of the coefficients in use in the models. I expect the sign of all age variables to be positive, given the gravitational modelling framework I will use. I expect the unemployment rate variable to propel households from the origin and dampen destination inflows. For the annual average pay variable, I expect the origin-based coefficients to be negative because a higher annual average pay is a positive and will dampen outgoing flows and I expect the opposite for the destination coefficient: higher annual average pay will attract households. For the median house value variable, I expect the origin coefficients to be positive because households will want to relocate to areas with a lower cost of living and I expect areas with a high cost of living, as evidenced by a higher median house value, to be viewed negatively and therefore areas with a high cost of living will dampen incoming flows. Wet and cold winters will propel households and dry and warm winters will attract households. Conversely, counties with a higher amenity score will dampen outgoing flows and attract incoming flows. Finally, I expect the distance between counties to decrease movement between the origin and the destination while I expect the adjacency and within-state coefficient to be positive.

Table 4.1: Source and hypothesized sign of variables

Variable Group	Variable Source	Variable Name	Hypothesized Coefficient Sign	
			Origin	Destination
Age	US Census Bureau	Less than 10	+	+
	US Census Bureau	Ages 10 - 19	+	+
	US Census Bureau	Ages 20 - 29	+	+
	US Census Bureau	Ages 30 - 39	+	+
	US Census Bureau	Ages 40 - 49	+	+
	US Census Bureau	Ages 50 - 59	+	+
	US Census Bureau	Ages 60 - 69	+	+
	US Census Bureau	Ages 70 - 79	+	+
	US Census Bureau	Ages 80 and older	+	+
Economic	Bureau of Labor Statistics	Unemployment Rate	+	-
	Bureau of Labor Statistics	Annual Average pay	-	+
	Census 1990, 2000, ACS 2008 - 2012	Median House Value	+	-
Amenity	NOAA	Inches of Precipitation	+	-
	NOAA	Heating Degree Days	+	-
	USDA	Amenities	-	+
Spatial	US Census Bureau	Distance	-	
	US Census Bureau	Adjacent	+	
	US Census Bureau	Same State	+	
	IRS, US Census Bureau	Accessibility	-	-

I expect the accessibility coefficients to be negative on account of the spatial distribution of the accessibility variable - higher accessibility measure in the eastern portions of the country. Households are generally moving west and the accessibility measure is less in the western portion of the country.

4.3 Model specification

The modelling framework in this chapter considers county origins I and county destinations J by each county's metropolitan and non-metropolitan status. This enables the comparison of the four types of household movement by origin push factor and destination pull factors.

This work includes origin factors (as opposed to just destination pull factors) because those have been found to be significant in determining out-migration (Clark and Ballard 1980). This consideration of movement is possible because of the work undertaken in chapter two wherein I estimated the explicit origins and destinations of all county-to-county household movement. The combination of 26-years of household movement data at a semi-fined grained spatial resolution with origin and destination explanatory variables is further enhanced by the specification for four different model types. This is done to accentuate the differences between moves to non-metropolitan and moves to metropolitan areas.

The modeling strategy is to expand upon the base gravity model:

$$T_{ij} = \frac{X_i X_j}{D_{ij}} \tag{4.2}$$

By including vectors of origin specific variables, destination specific variables, and pair-wise specific variables to fit four models per year for a total of 104 models. The observations for each model are selected to reflect the four different types of movement per year and with 26-years of household movement data a total of 104 models will be fit:

1. 26 models featuring non-metropolitan-to-non-metropolitan movement.

2. 26 models featuring non-metropolitan-to-metropolitan movement.

3. 26 models featuring metropolitan-to-non-metropolitan movement.

4. 26 models featuring metropolitan-to-metropolitan movement.

Models one and three will describe how place-based characteristics contribute to non-metropolitan settlement and counter-urbanization. Models two and four will illustrate how different determinants promote metropolitan settlement. The dependent variable is assumed to follow a Poisson distribution (Flowerdew and Aitkin 1982). Counts of households moving from origin county i to destination county j are linked to a linear combination of logarithmically transformed origin and destination specific variables:

$$\lambda_{ij} = exp(\beta_{xi} log(X_i) + \beta_{xj} log(X_j) - \beta_{D_{ij}} log(D_{ij})) \tag{4.3}$$

In the reduced-form model in equation 4.3, X_i is the vector of origin specific variables, X_j is the vector of destination specific variables, and D_{ij} is a vector of variables describing the spatial relationship between each origin i and each destination j. All variables, except for the Boolean variables, are log-transformed to accommodate the logarithmic-link specification necessary for modelling data with a Poisson distribution. Given the distribution of the dependent variables, several large county-to-county flows and many small county-to-county flows, I will account for this over-dispersion by using a quasi-Poisson model. Another option would be to use a negative binomial model. The Poisson model assumes the variance of the dependent variable is equal to the conditional mean of the dependent variable. The quasi-Poisson model treats the variance of the dependent variable as a linear function of the mean and the negative binomial model treats the variance as a quadratic function of the mean. Because of this treatment, large and small counts are weighted differently in the quasi-Poisson model versus the negative binomial model. In a negative binomial model, smaller counts are given more weight than in a quasi-Poisson model whereas weighting is constant in a quasi-Poisson model (Ver Hoef and Boveng 2007). For this modelling exercise, using the quasi-Poisson specification will give equal weight to all county-to-county flows which in turn will generate coefficients within a more compressed range of values speaking to the push and pull factors of each variable by movement type.

4.4 Results

The 104 regression models were fit using R 4.04. The goodness-of-fit of these models was judged by the percent-of-deviance explained. The percent of deviance explained ranges from 68-percent to 88-percent. In general, the non-metropolitan-to-non-metropolitan and metropolitan-to-metropolitan models perform better than the non-metropolitan-to-metropolitan and metropolitan-to-non-metropolitan models. The non-metropolitan-to-non-metropolitan and metropolitan-to-metropolitan models explaining an average of 76-percent and 87-percent

of the deviance, respectively. The non-metropolitan-to-metropolitan and the metropolitan-to-non-metropolitan models explain greater deviance over time. These 104 regression models generated 3,640 coefficients (35 covariates * 4 models per year * 26-years) and all but 174 of these coefficients, 95.2-percent, are significant at the 0.05 level. This is to be expected given the large sample size in each movement type. The non-metropolitan-to-non-metropolitan models feature approximately 5.5 million records per year, the non-metropolitan-to-metropolitan and metropolitan-to-non-metropolitan models each feature approximately 1.8M records per year, and the metropolitan-to-metropolitan models each feature approximately 0.6M records per year. There are few patterns within the non-significant coefficients. The spatial structure variables are always significant. As are the coefficients for children and teenagers ages 10 through 19 and adults ages 20 through 29. Over half of the non-significant coefficients are in the other age variables. All non-significant coefficients feature a value near zero. Approximately 32-percent of the non-significant coefficients come from the metropolitan-to-metropolitan models with years 1991 through 1995 and 2008 through 2010 featuring the most non-significant coefficients. The year with the models featuring the most non-significant coefficients, 11, is 1996. The year with the least is 2007 with 2 non-significant coefficients.

I have a prepared a line graph illustrating each coefficient's trend over time for each move type. In addition, I have included the 95-percent confidence interval for each coefficient. Due to the large sample in each model, the confidence intervals are small and not always discernable in the visualizations. The distance, spatial structure, economic, and amenity variables are colored by movement type while the age structure coefficients are colored by age group. Movement types with the same origin and destination categories are colored with darker hues while movement types with different origin and destination categories are colored with lighter hues. Moves originating in non-metropolitan counties are colored purple and moves originating in metropolitan counties are colored shades of grey. Non-significant coefficients are identified by a green dot in the spatial structure, economic, and amenity variable graphics and a black dot in the age group graphics.

The choice to differentiate by movement type or age group is done to facilitate comparison within and across determinants. These coefficients were generated from a generalized linear model. Given the specification of the model, values of a coefficient less than zero indicate that a coefficient is dampening outgoing or incoming migration flows and values of a coefficient greater than zero indicate that a coefficient is increasing outgoing or incoming flows. A negative coefficient indicates that the effect on household movement is divisive and positive coefficients indicate the effective is multiplicative. I will first describe the trends present in the distance and spatial structure variables followed by the age structure variables, the economic variables, and finally the amenity variables.

4.4.1 Distance and spatial structure

The visualizations of the distance and spatial structure coefficients feature coefficients colored by movement type. Figure 4.12 on page 156 features the coefficients of distance, graphic A, the county adjacency indicator, graphic B, and the within-state indicator, graphic C. Note that the y-axis is graduated differently in each graphic to better emphasize the trends present in each coefficient. Because the distances between counties are fixed and the county metropolitan classifications are fixed to 1990, the regression coefficients are only influenced by the numbers of households moving between the various sets of origins and destinations in any given year.

For all move types, distance is a deterrent to movement between counties as all coefficients are negative during the 1991 through 2015 period. For all move types, beginning in 2012, the distance coefficient grew less negative indicating that households were moving further. The coefficients showcasing the effect of nearness and borders on movement are presented in graphics B and C in Figure 4.12. Graphic B shows the coefficient for directly adjacent origin and destination counties and graphic C shows the coefficient for origin and destination counties within the same state. The adjacency coefficient, for all models, is quite large indicating that regardless of movement type, a move to an adjacent county is preferential. Especially so for non-metropolitan-to-non-metropolitan movement. The coefficient for households in non-

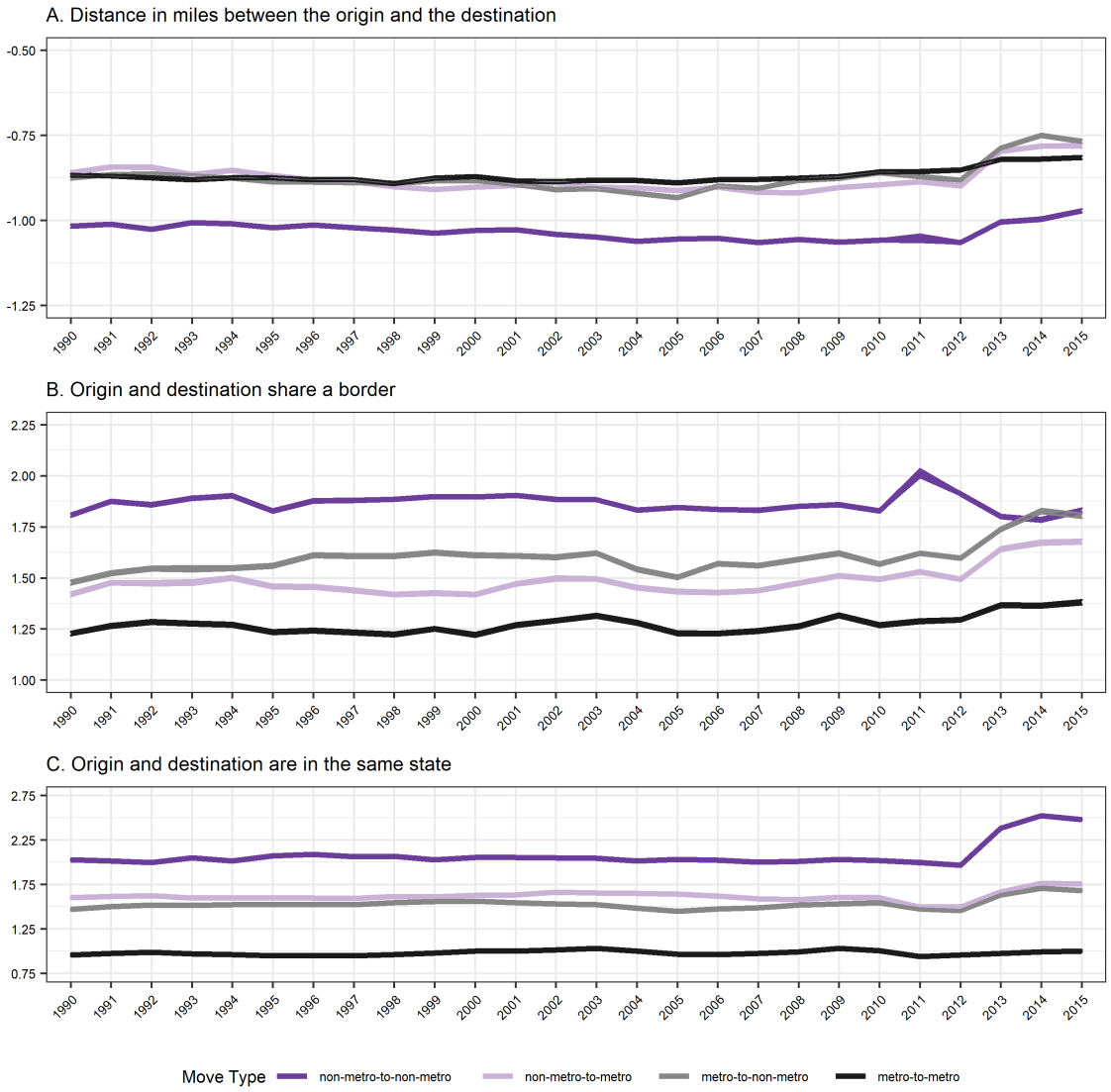


Figure 4.12: Coefficients of distance and nearness by move type, 1990-2015

metropolitan counties moving to directly adjacent non-metropolitan counties peaked in 2011 indicating that directly adjacent destinations were less of a draw. For non-metropolitan-to-metropolitan and metropolitan-to-non-metropolitan moves, adjacent counties generated more movement. Moves between adjacent metropolitan counties trended upwards during the study period. In general, the adjacency coefficients for metropolitan-to-metropolitan moves and metropolitan-to-non-metropolitan moves are more similar to each other in any given year, than when compared to the non-metropolitan-to-non-metropolitan coefficients and the non-metropolitan-to-metropolitan coefficients. This is in part due to the adjacency variable acting as a pseudo-proxy for residential mobility. For moves that originate in a metropolitan area that then terminate in a non-metropolitan area, these are households moving to non-metropolitan counties adjacent to metropolitan counties, possibly reflecting a desire to stay near known employment centers and social networks. For moves originating in non-metropolitan counties and terminating in adjacent metropolitan counties is still preferential, but less so when compared to metropolitan-to-non-metropolitan moves, indicating that non-metropolitan households move further to metropolitan destinations.

The positive coefficients in graphic C in Figure 4.12 illustrate how within-state destinations draw additional households across all movement types. State borders matter for internal migration. The movement types exhibit minor year-over-year change during the 1990 through 2011 period indicating a consistent preference for within-state destinations. The smallest coefficients are seen in metropolitan-to-metropolitan movement while the largest coefficients are seen in non-metropolitan-to-non-metropolitan movement. For metropolitan-to-metropolitan movement, this means that within-state metropolitan origins and destinations, while still generating additional flows of households, are not as generative as movement between non-metropolitan counties. Non-metropolitan origins and non-metropolitan destinations in the same state generate more relative movement between them when compared to metropolitan origins and metropolitan destinations in the same state. The similar coefficients for non-metropolitan-to-metropolitan and metropolitan-to-non-metropolitan within-state movement suggest similar migration streams between the two movement types. Starting in 2013 each

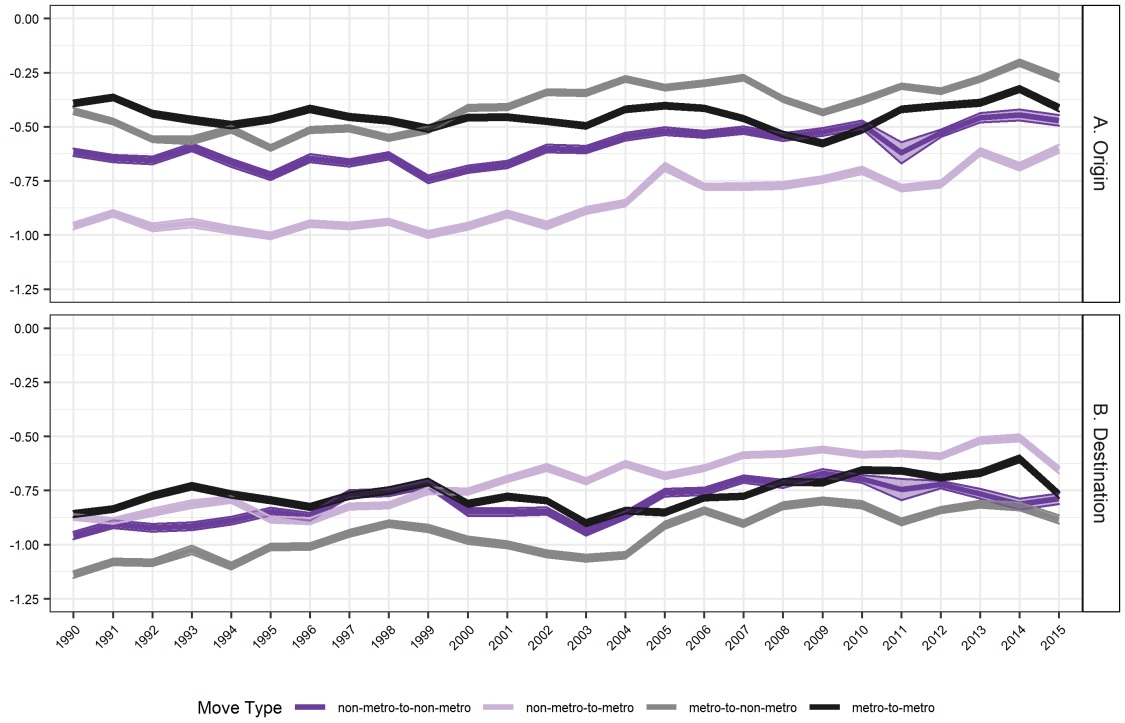


Figure 4.13: Coefficients of accessibility by move type, 1990-2015

movement type featured a greater affinity for an in-state destination. It is worth nothing that some of the variation in the coefficients is due to a change in IRS data tabulation procedures (DeWaard et al. 2020b; Pierce 2015).

The origin and destination accessibility coefficients are presented in Figure 4.13 on page 158. The accessibility coefficients are negative for all years indicating that increases in both the origin county's accessibility and the destination county's accessibility decrease the expected count of flows moving between an origin and a destination. At first, this seems counterintuitive, but reviewing the accessibility measure will illustrate the mechanism driving these values. The accessibility measure is the sum total of the number of households-per-mile that would have to move to reach a focal county. As this measure is a sum of ratios, increasing the numerator (more households) or decreasing the denominator (shorter distances) increases the accessibility measure. More centrally located counties have higher accessibility measures as

do counties in the eastern half of the US. The counties with the largest accessibility measures are small counties close to large population centers. A heavily populated county will have a lower accessibility measure than its sparsely populated neighbor. Therefore, this relationship is inversely proportional with population size: as the population of a focal county increases, the accessibility measure of the county decreases and therefore less populated counties are going to have greater accessibility measures. In this sense, the origin and destination accessibility measures showcase a county's emissive potential and attractive potential as a function of the county's population size and relative spatial position. A county's population size and relative spatial position matters more so for destinations, as indicated by the more negative values, than for origins, regardless of move type. Phrased differently, a county with a large accessibility measure (a less populated county), is going to dampen incoming flows more so than it will dampen outgoing flows. This is especially so for metropolitan-to-metropolitan movement. The similarity of the coefficients over time reflects the degree to which distance and metropolitan categories are held constant.

4.4.2 Age structure

The nine age structure variables show how different age groups promote, or dampen, outgoing or incoming household flows, respectively. While gravitational theories of movement suggest that increases in origin or destination mass should increase interaction between the origin and destination, and should therefore be positive, the coefficients of the age structure variables indicate different trends. I initially hypothesized that increases in any one age group would promote an increase in the number of inbound or outbound households and those hypotheses would hold true over time. As will be shown, this is not the case. To communicate the trends in the coefficients in each of the nine age groups, I have prepared three separate visualizations with each visualization featuring eight separate graphics and each plot featuring coefficients for three age groups. The coefficients for the youngest age group in each plot in each visualization are colored blue, the coefficients for the oldest age group in each plot in each visualization are colored red, and the coefficients for the age group in the

middle in each plot in each visualization are colored green. The rows of each visualization pertain to the movement type and the columns of each visualization pertain to the origin and destination coefficients. A black dot indicates a non-significant coefficient. The graphics for each movement type feature the same y-axis scale to help with comparison across age groups. The dependent variable in every model is the number of households moving from one county to another in any given year. The age variables are the number of people in a specific age group. The coefficients of the number of people in each of the nine age groups represent the generative or attractive potential of the mass of people in these age groups. For nearly all age groups across all movement types in all years during the 1990 through 2015 period, the origin coefficient features a different value than the destination coefficient, regardless of movement type. This indicates that each age groups influences outbound household migration and inbound household migration differently. Finally, what is frequently seen in each age group - but not always - is that for coefficients across all movement types exhibit one pattern through the 1990s, an inflection point at some time in the early 2000s, and possibly another inflection point in the late 2000s, shortly after the Great Recession. These periods roughly correspond to periods of positive non-metropolitan net-migration rates. Before discussing the generative and attractive potential of each age group, I will describe a counterfactual scenario illustrating the utility of disaggregating by age.

The counterfactual scenario

One way to ascertain the usefulness of the proposed modelling tactic of disaggregating the origin and destination population by age groups is to examine an alternative approach: combining all nine age groups into a single category or ignoring the metropolitan status of the origin and destination. Figure 4.14 on page 162, showcases the results from an additional set of regression models using a combined population count for the age groups. Five regression models with the same specification were fit each year. The models take the same general form as previously described with one important distinction: each model, in addition to the spatial structure covariates (distance, adjacency, within-state, and accessibility)

and the economic and amenity variables, features only a count of the population at the origin and the count of the population at the destination: all nine age groups were combined into a count of the total population. The five types of regression models pertain to different forms of movement. The non-metropolitan-to-non-metropolitan model coefficients are shaded dark purple, the non-metropolitan-to-metropolitan model coefficients are shaded light purple, the metropolitan-to-non-metropolitan model coefficients are shaded light grey, and the metropolitan-to-metropolitan model coefficients are shaded dark grey. An all-movement model that ignores the metropolitan status of the origin and destination counties is included, and it is shaded red. The results show that with some minor variation and a peak in 2011 in the metropolitan and non-metropolitan coefficients the emissive potential of the total population ranges between approximately 0.75 and 1.00. Because these coefficients are always greater than zero during the study period, the all-age groups combined population does induce additional outgoing households and draw additional flows. The disaggregation of the population into nine age groups shows how different age groups push and pull households at different rates with variation over time.

Age groups 9 and younger, 10-19, and 20-29

The first visualization with the age structure coefficients, Figure 4.15 on page 163, features the effect of young children ages 9 and younger, children and teenagers ages 10 through 19, and young adults ages 20 through 29, on county-to-county movement by movement type. Across all movement types and over time, the numbers of children nine and younger (coefficients colored blue) generate differentials in outgoing and incoming migration. From a lifecourse perspective, this would suggest that at different points in time, households with younger children are more likely to move, these households are less rooted, while households with older children and teenagers are more rooted in general. Positive destination coefficients indicate that non-metropolitan and metropolitan destinations with children ages 9 and younger draw additional households. This would mean that households with young children are going to move to locations where there are already young children (i.e., there are

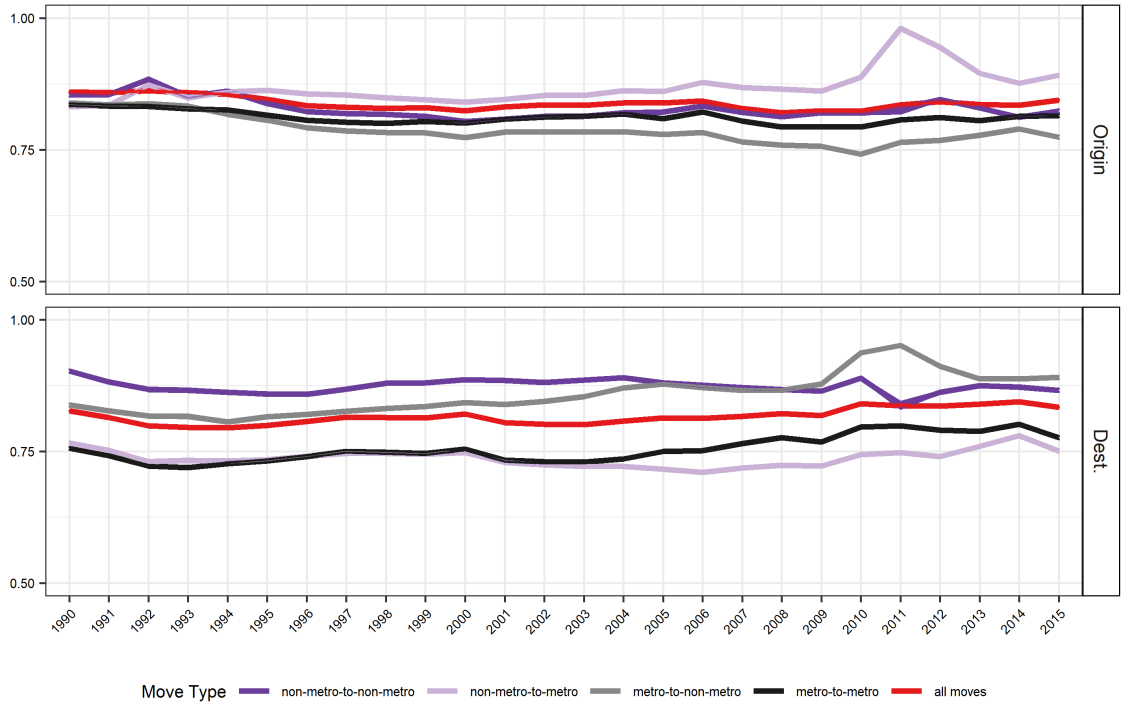


Figure 4.14: Coefficients of the all ages population by movement type, 1990-2015

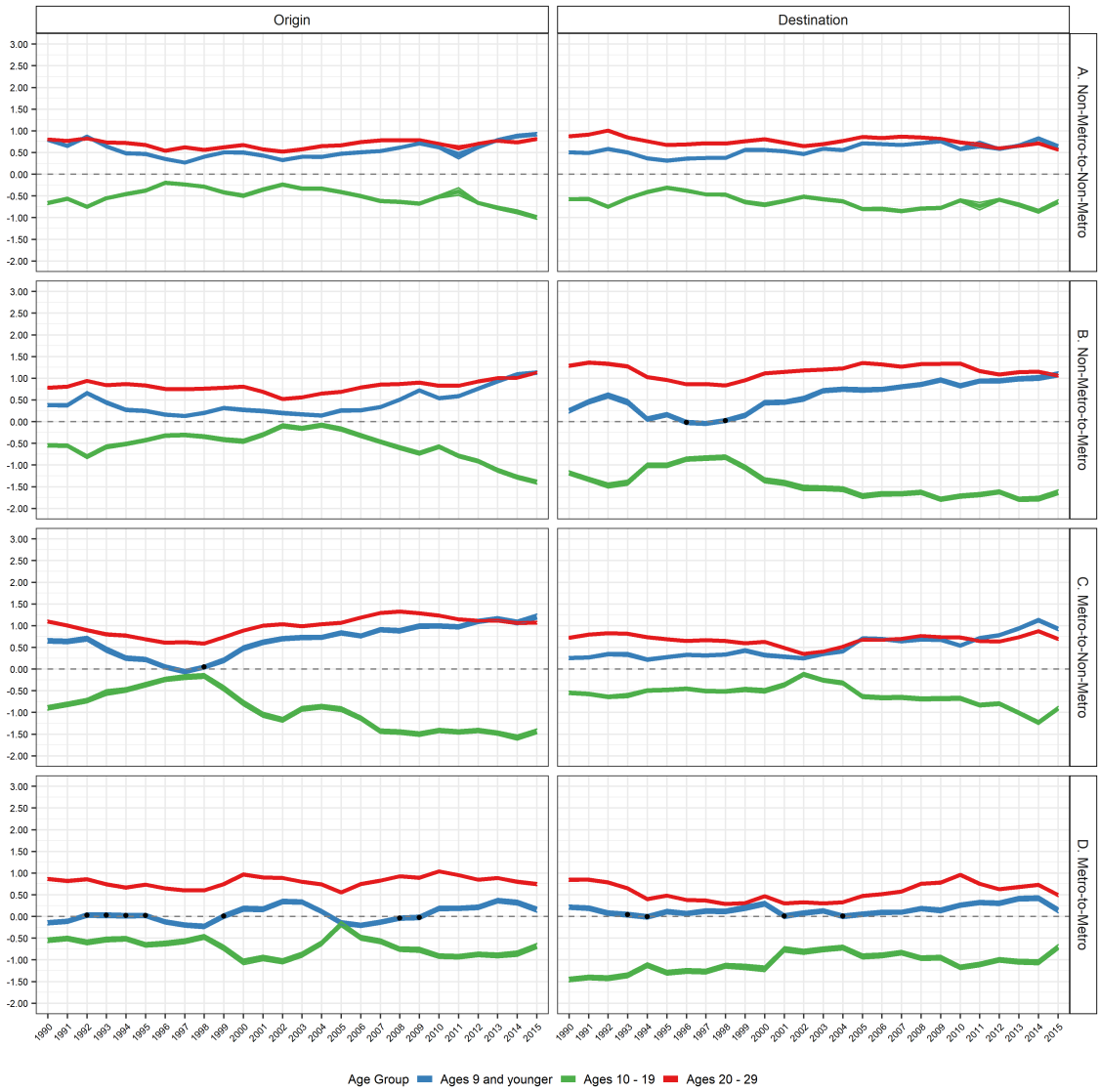


Figure 4.15: Coefficients off age groups nine and younger, 10-19, and 20-29, by movement type, 1990-2015

schools in the destination). When the coefficient for this age groups dips below zero, both the counts of outgoing and incoming households are deflated. For origins, for years with negative coefficients, late 1990s through mid-2000s, this population dampened outgoing flows. For destinations, for years with negative coefficients, the number of children ages 9 and younger at the destination did not draw additional households. For metropolitan-to-metropolitan migration during the early 2000s, this suggests that households were less inclined to move to metropolitan counties with numbers of children. The generally positive values of the coefficients of children ages 9 and younger is in line with previous studies indicating that children in this age group are mobile in that they are tied to their parents who migrate.

Across all movement types and over time, the numbers of children and teenagers between 10 and 19 (coefficients colored green) dampen both outgoing and incoming flows. The coefficients for the number of children and teenagers ages 10 through 19, as seen in row B of Figure 4.15, are almost always negative. This age group does not induce the migration of additional households for origin counties, nor does it draw additional households for destination counties. Rather, the number of children and teenagers ages 10 through 19 suppress both outbound and inbound flows. This suggests that households with children and teenagers in this age group are more rooted and less of a draw.

The impact on household migration of the number of adults in the 20-29 age group (coefficients colored red) is quite different, the coefficients for this age group are always positive. Members of this age group are themselves filing taxes and are no longer dependents on tax returns. Historically, this age group is very mobile as adults in this age group are typically moving for college or employment. The potential of this age group to increase outbound migration decreases from higher potentials in the early 1990s, reaches a low point in the early 2000s, and in the mid-2000s through 2015, the effect is multiplicative. Young adults in metropolitan areas generate greater outbound flows and greater inbound flows.

Figure 4.16: Coefficients of age groups 30-39, 40-49, and 50-59, by movement type, 1990-2015

Age groups 30-39, 40-49, and 50-59

The second graph with the age structure coefficients, Figure 4.16 on page 165, features the effect of adults in their 30s, 40s, and 50s, on county-to-county migration by the four movement types. The coefficients of these three age groups feature the greatest range of values indicating larger variability in a particular age group's ability to generate additional outgoing households and draw additional incoming households. The coefficients of the 30-39 age group (colored blue), generally begin decreasing in the late 1990s, reach their lowest values around 2000, and either maintain that value or begin a slow increase. For all four types of movement the coefficients for 30-39 age group show two distinct periods of movement: during the 1990s this group demonstrated a steady ability to motivate households to leave an origin and draw additional households to the destination. After 2000, this group's ability to propel and draw households decreased.

The coefficients in the 40-to-49 age group (colored green) are nearly always positive, with minor exceptions, indicating that this group has been both propelled and attracted additional households. The ability of this age group to propel and attract non-metropolitan households reached its peak around 2000. For households leaving metropolitan counties or arriving in metropolitan counties, the 40-to-49 age group reached peak propulsive and attractive potential five to seven years later in the late 2000s. For metropolitan-to-metropolitan movement, this age group consistently propelled and attracted households.

The final set of coefficients in Figure 4.16 are for the 50-to-59 age group (colored red). In general, across the four movement types, this age group shows greater potentials after 2000 indicating that people ages 50-59 both propelled households and attracted households in greater numbers. This trend is seen in all four types of movement in both the origin and the destination coefficient suggest. For non-metropolitan-to-non-metropolitan movement, this age group deflated outgoing and incoming migration from 1990 through 2000 and beginning in the early 2000s, propelled and attracted households. Across all movement types, a consistent trend in the effect of people in their 30s, 40s, and 50s is pronounced difference before 2000

and after 2000. This suggests a period effect in the coefficients.

Age groups 60-69, 70-79, and 80 and older

The final set of age group coefficients is on display in Figure 4.17 on page 168. These coefficients feature some of the largest and the smallest range of values. The 60-69 age group, colored blue, features the largest absolute values indicating that of the three age groups in Figure 4.17, the 60-69 age group features the greatest propulsive and emissive potential. For non-metropolitan-to-non-metropolitan movement, the coefficients of the 60-69 age group are almost always positive indicating that the age group's ability to propel household from and attract households to non-metropolitan areas is quite strong. The attractive potential is greater than the emissive potential. For non-metropolitan-to-metropolitan movement, the 60-69 age group featured a higher emissive potential in the 1990s with a dip around 2000. The attractive potential of the group reached its peak in 1999 for non-metropolitan-to-metropolitan movement. For metropolitan-to-non-metropolitan movement, the ability of the group to propel additional households decreased over time, experienced a slight increase in the early 2000s, and then decreased again. The 60-69 age group's ability to attract households in metropolitan areas to non-metropolitan areas is always strong, especially in the late 1990s. This is consistent with other findings showing increased mobility in retirement age (Bernard et al. 2014). Three different periods of coefficients are seen in the metropolitan-to-metropolitan movement for the 60-69 age group: greater emissive potential in the 1990s, decreases in the 2000s, and increases in the mid-2000s.

The 70-79 age group, colored green, features several trends. For non-metropolitan-to-non-metropolitan, the coefficients of this age groups are almost always negative, indicating that this age group does not propel or attract additional households from and to non-metropolitan counties. The same is generally true for this age group for non-metropolitan-to-metropolitan movement. However, there is an increase in the 70-79 age group's ability to attract households moving from non-metropolitan counties to metropolitan counties. The 70-79 age group gradually increases in its ability to propel households from metropolitan counties and de-

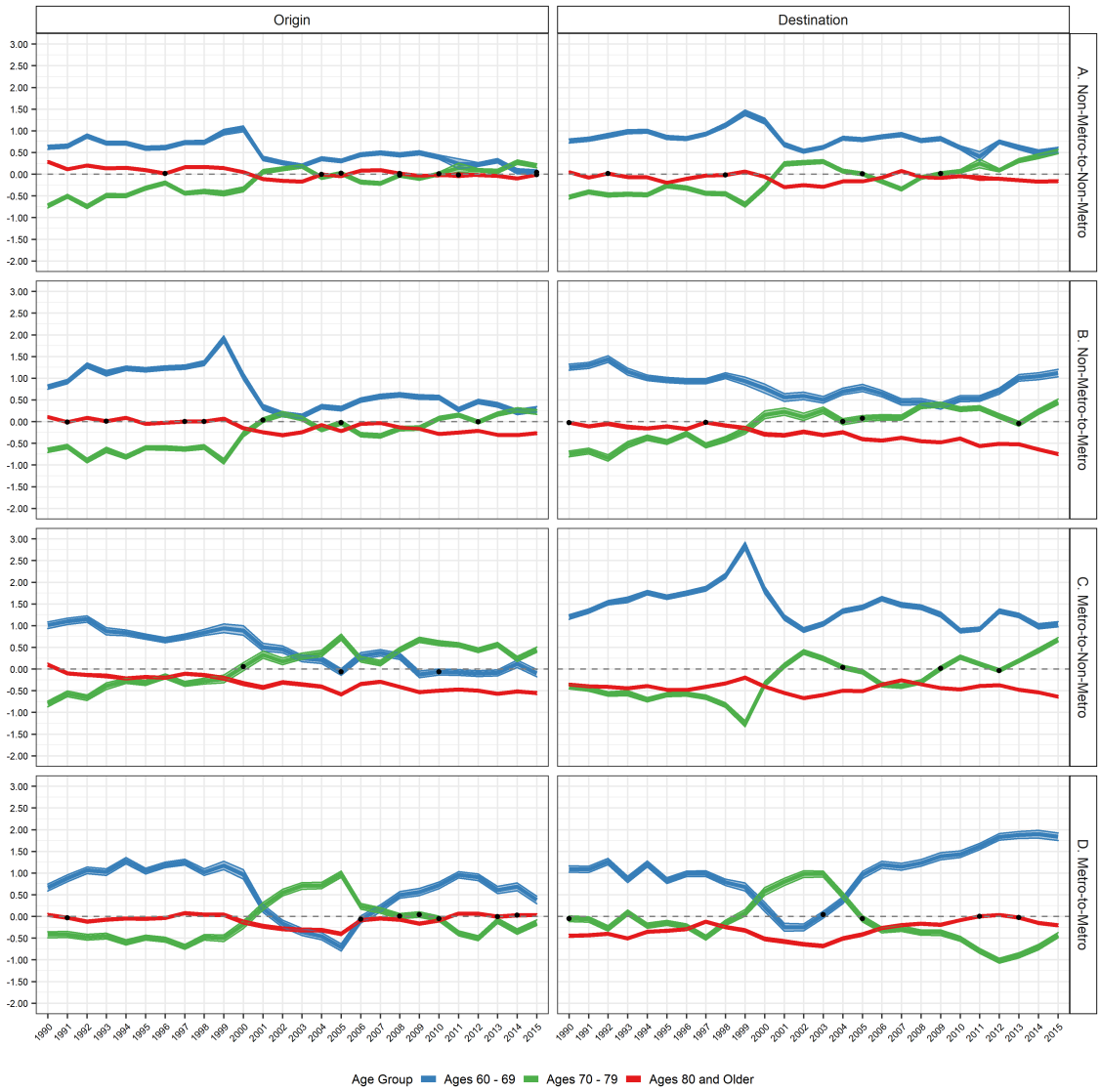


Figure 4.17: Coefficients of age groups 60-69, 70-79, and 80 and older

creases in its ability to attract households to non-metropolitan destinations. The 70-79 age group's emissive and attractive potentials for metropolitan-to-metropolitan movement both increased beginning in the late 1990s and began decreasing again in the mid-2000s.

Finally, the 80 and older age group coefficients, colored red, shows the least variation with most coefficients near zero. This is especially true with non-metropolitan-to-non-metropolitan movement. The 80 and older age group does not induce the migration of additional households from non-metropolitan areas while over time, the 80 and older age group decreased the count of households migrating into metropolitan counties. The 80 and older age group exhibits little variation in its ability to propel households from metropolitan counties to non-metropolitan counties. This age group does not feature a strong attractive potential for households moving to non-metropolitan counties from metropolitan counties. The 80 and older age group is generally consistent in its ability propel houses from metropolitan counties to other metropolitan counties, though its ability to attract households is limited.

Concluding this section, some age groups are more consistent than others in their emissive and attractive potential. The coefficients for children and teenagers ages 10-19 are almost always negative across all movement types indicating that this age group decreases both outgoing flows and incoming flows. The 20-29 age group is always positive indicating that this age group both propels outgoing household and attracts households. The 30-39 age group more has a greater emissive potential in the 1990s than in the 2000s and later. The 40-49 age group does attract households to metropolitan counties from both non-metropolitan counties and other metropolitan counties. In general, the 40-49 age group both propels and attracts households across movement types. The 50-59 age group does draw an increase in households moving from metropolitan counties to other metropolitan counties. The 60-69 age group always motivates and attracts households to move between non-metropolitan counties and this age group in general propels and attracts households, especially households moving from metropolitan to non-metropolitan counties. The 70-79 age group consistently dampens flows between non-metropolitan counties and flows emanating from metropolitan

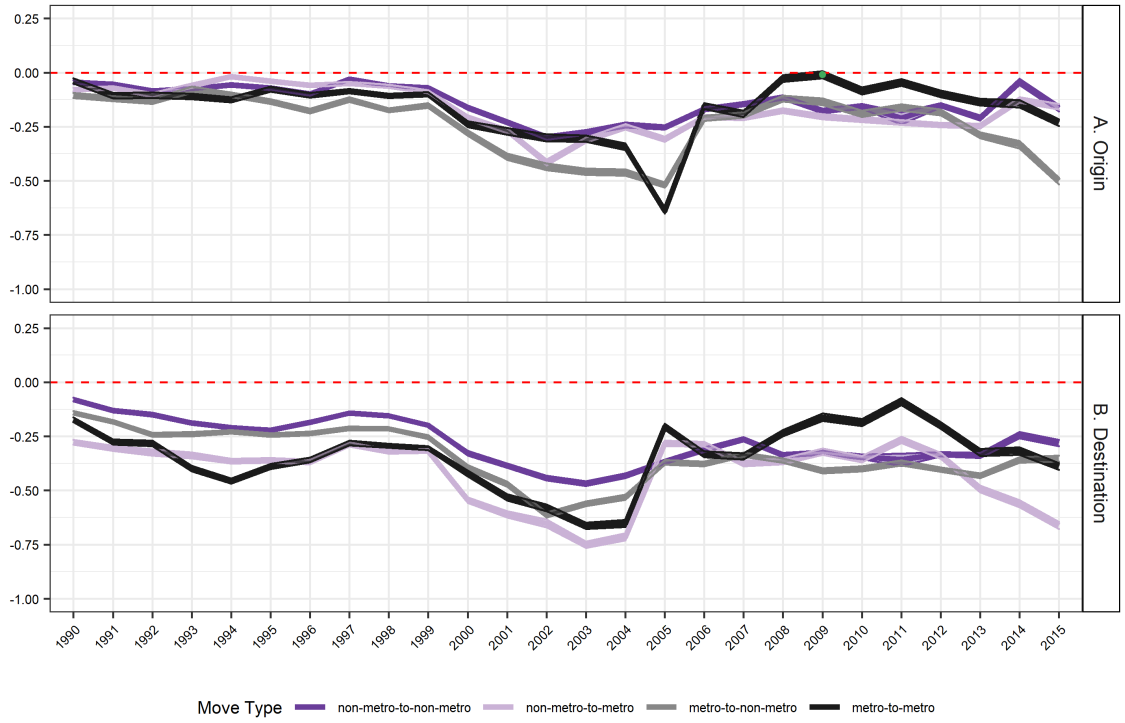


Figure 4.18: Coefficients of the unemployment rate, 1990-2015

counties. The 80 and older age group dampens flows emanating from metropolitan counties to other metropolitan counties.

4.4.3 Unemployment, annual average pay, and median house value

Figure 4.18 on page 170 features the coefficients of the unemployment rate visualized by the four movement types. The origin and destination unemployment rate coefficients are always negative across all four movement types indicating that unemployment decreases outgoing flows and decreases incoming flows. It makes intuitive sense that the unemployment rate at the destination dampens the flow of incoming households: a high unemployment rate means that jobs are more scare and a high unemployment rate at the destination is indicative of a distressed local area economy, a negative destination quality. The mechanism presumed to

be behind how unemployment rates dampen outgoing flows is less intuitive. Higher rates of unemployment indicate an economically depressed local area. Households in an economically depressed area are already strained and might not have the resources to move to other areas. In the late 1990s through the mid-2000s, the degree to which unemployment dampened outgoing flows grew in strength, but more so for flows originating in metropolitan areas. By 2006, the degree of dampening lessened slightly and by 2008, the degree to which the unemployment rate dampened outgoing flows decreased. The unemployment rate in non-metropolitan counties dampens flows of outgoing households differently than in metropolitan households. This suggests that households in non-metropolitan counties respond to unemployment rates differently than households in metropolitan counties.

The effect of the destination rate of unemployment follows a pattern similar to the origin rate of unemployment, just with a greater degree of dampening. A general decrease from 1990 through 2003 and then increasing from 2004, but never reaching the less negative values seen in 1990. The general trend over time is that metropolitan and non-metropolitan unemployment rates became more of a destination deterrent. This reflects the fact that destinations with higher unemployment rates are not perceived as attractively. The largest change is seen in metropolitan-to-metropolitan movement, indicating that households moving between metropolitan counties are reacting different to unemployment rates than households undertaking different types of movement. If households are moving between metropolitan areas because of jobs, it makes sense that households are going to move to metropolitan areas with lower rates of unemployment. The unemployment coefficient for non-metropolitan-to-non-metropolitan movement grew more negative throughout the 1990s, with a slight uptick in 1996 and 1997, and reached its minimum in 2003. The coefficient increased from 2004 through 2007 where it plateaued. The shape of the curve for metropolitan-to-non-metropolitan unemployment rate coefficient is similar, but always more negative. This means that metropolitan households and non-metropolitan households react differently to non-metropolitan unemployment with households in metropolitan areas finding it less of a destination deterrent (the 1990s) and more of a deterrent (2005 through 2015)

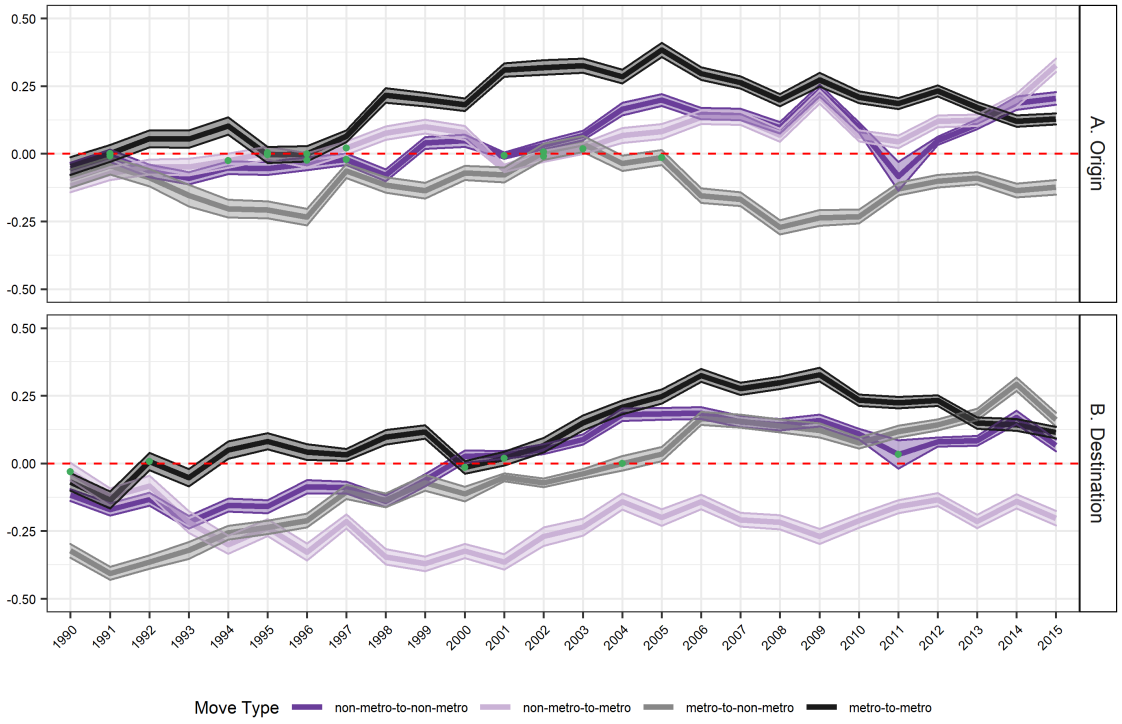


Figure 4.19: Coefficients of annual average pay, 1990-2015

The coefficients of the annual average pay are shown in Figure 4.19 on page 172 For moves originating in metropolitan counties and non-metropolitan counties, the ability of a county's annual average pay to promote the number of outgoing households increased from 1990 through the early and mid-2000s, but with some minor decreases in between. The difference is when the annual average pay of the county started decreasing in its ability to promote outgoing flows. For flows originating in metropolitan counties, this occurred in the early 2000s and for flows originating in non-metropolitan counties this occurred in 2009. For metropolitan-to-non-metropolitan movement starting in 2003 and lasting until 2008, the annual average pay in the origin county dampened outgoing flows until it became increasing again. This period roughly corresponds to net positive non-metropolitan migration rates.

The annual average pay coefficients for the destination show generally upward trends with the exception of non-metropolitan-to-metropolitan movement. For non-metropolitan-to-metropolitan movement, destination pay dampens incoming flows, reflecting the differences in annual average pay in metropolitan and non-metropolitan counties. By 2004, the coefficients for non-metropolitan-to-metropolitan movement were steady suggesting that annual average pay was less differentiated between metropolitan counties. For non-metropolitan-to-non-metropolitan and metropolitan-to-non-metropolitan movement, the coefficients were positive, and plateaued after 2003. Non-metropolitan-to-non-metropolitan and metropolitan-to-non-metropolitan movement was undertaken by households seeking to maximize pay gains by moving between and to non-metropolitan counties. Starting in 2000, the destination coefficient for annual average pay increased, except for non-metropolitan-to-metropolitan movement, indicating that differences in annual average pay was driving household migration to greater degrees.

The final set of economic coefficients, median house value, are featured in Figure 4.20 on page 174. The median household variable is a measure of affordability and cost of living. I initially expected the origin coefficients to be positive and the destination coefficients to be negative. For non-metropolitan-to-non-metropolitan movement, the coefficients are positive in the 1990s, negative through the late 2000s, and positive again after 2007. Both

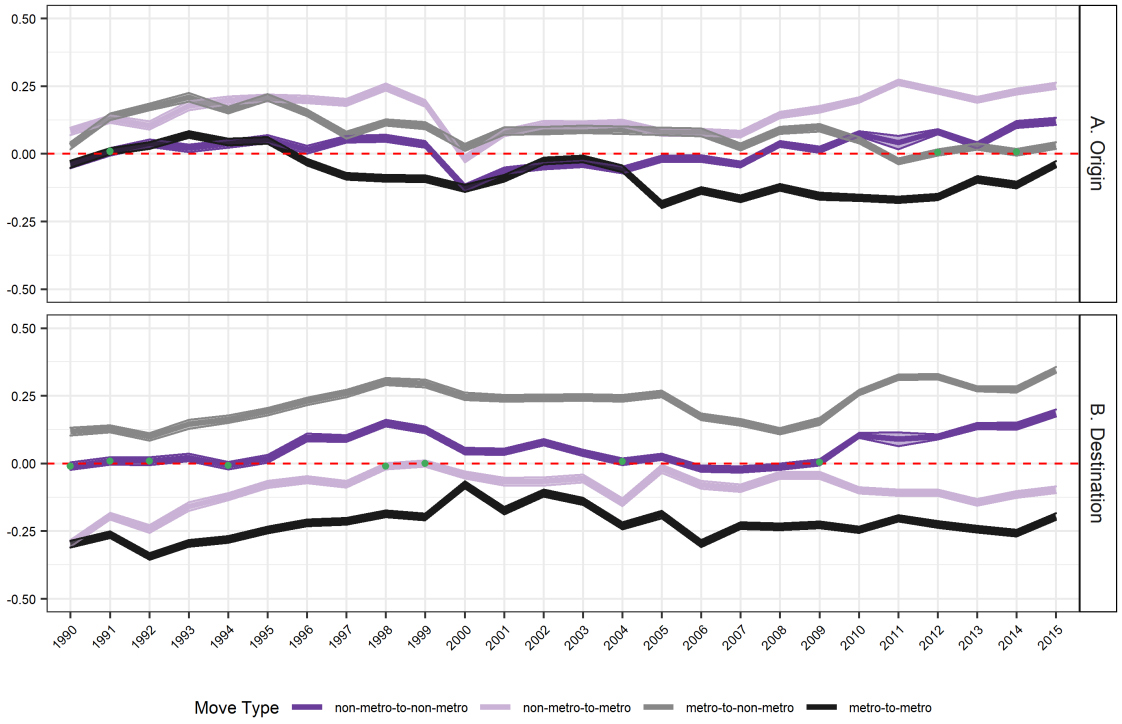


Figure 4.20: Coefficients of median house value, 1990-2015

the non-metropolitan-to-metropolitan and metropolitan-to-non-metropolitan curves are similarly shaped, just with different magnitudes. The most frequently negative curve is the metropolitan-to-metropolitan curve indicating that high home values dampen flows more so between metropolitan areas. This is most likely due to the cost of selling an expensive house and finding a similar house in a different metropolitan county. The high cost of living in metropolitan counties is a push factor for households moving from metropolitan to non-metropolitan counties. The destination coefficients are split by destination type. Households are attracted to non-metropolitan destinations with a higher cost of living. Households moving to metropolitan counties are dissuaded by higher home values. Juxtaposing the annual average pay coefficient and the median house value coefficients for metropolitan-to-metropolitan movement, households are drawn to lower costs of living and greater rates of pay.

4.4.4 Precipitation, heating degree days, and outdoor amenities

The final three coefficients in the discussion of the determinants are the total inches of rain in Winter and the number of heating degree days in Winter. Figure 4.21 on page 176, features the coefficients of the total number of inches of rain in winter, 1990 through 2015. The origin coefficient is generally negative, indicating that wet winters depress outgoing flows. This is a counterintuitive finding, and the explanation has to do with the proportion of moves to directly adjacent counties. In general, the number of inches of rain between any two adjacent counties is going to be very similar and the degree to which precipitation in winter dampens outgoing flows is negligible. The destination coefficients are also mostly negative and within a narrow range, indicating that wet winters moderately depress incoming flows. This is particularly pronounced in metropolitan-to-metropolitan movement and less so in the other types of movement. In 2002, the destination coefficients for metropolitan-to-non-metropolitan movement increased to above 0.0 suggesting that wet winters were a draw for households leaving metropolitan counties. This most likely reflects amenity driven migration. After 2009, the destination coefficients for the total number of inches of rain were

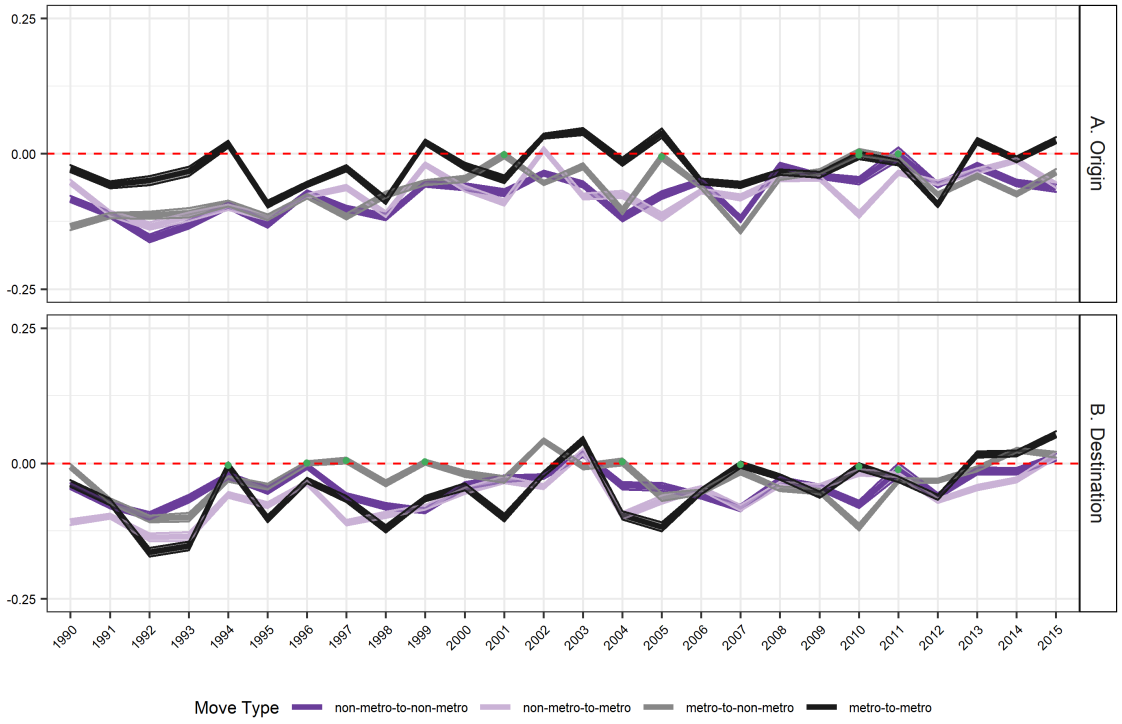


Figure 4.21: Coefficients of the total number of inches of precipitation in winter, 1990-2015

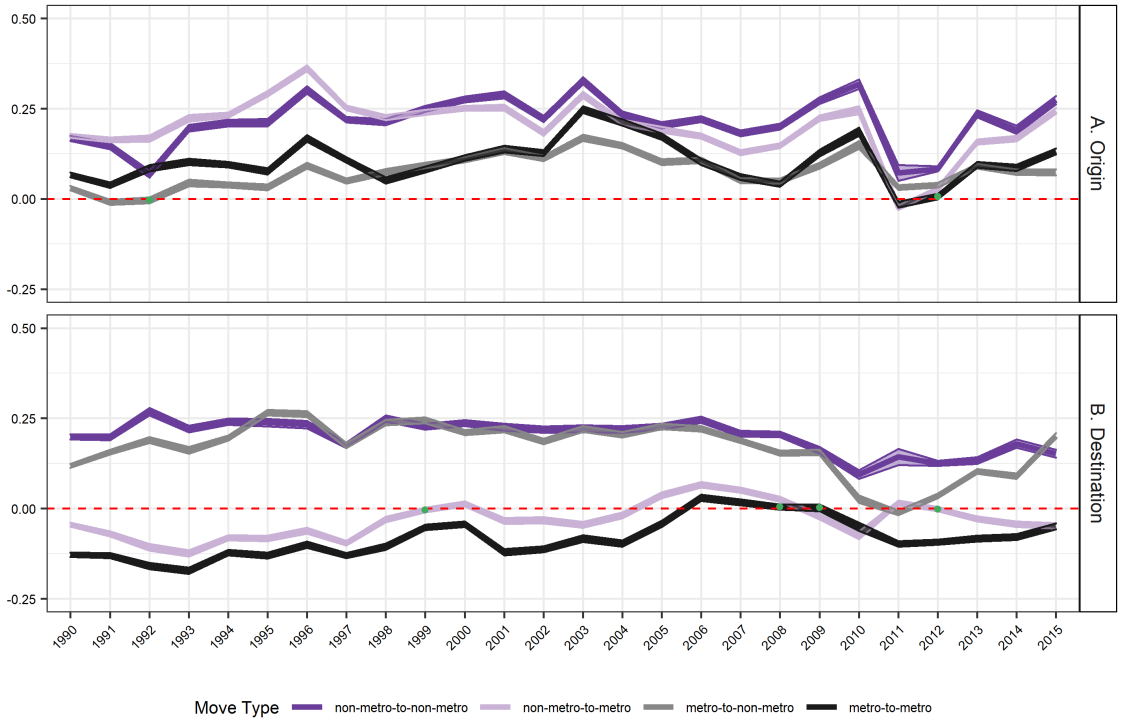


Figure 4.22: Coefficients of total number of heating degree days in winter, 1990-2015

increasing, though still generally negative.

. Figure 4.22 on page 177 features the coefficients of the total number of heating degree days in winter for 1990 through 2015. The origin-based heating degree day coefficients show that across all movement types, colder areas propel households. A temporary, but noticeable decrease was seen in 2011. Colder non-metropolitan counties propel households at greater rates than metropolitan counties as indicated by the greater coefficients. Throughout the 1990s, the coefficients of the four movement types were distinct, indicating households in metropolitan and non-metropolitan counties reacted to cold weather differently and this was influenced by the type of destination. By 2000, the coefficients for non-metropolitan counties converged and the coefficients for metropolitan counties converged indicating that destination type was less of a factor and the migration response was more so indicative of the immediacy of the weather.

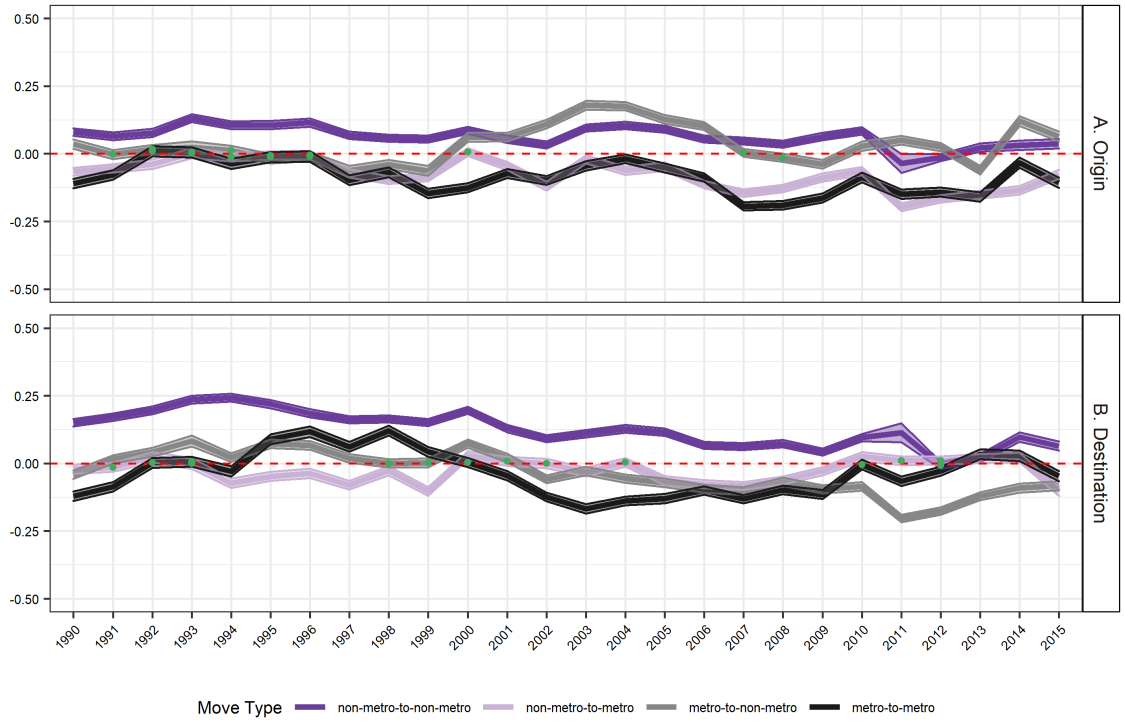


Figure 4.23: Coefficients of the amenity scale, 1990-2015

Cold weather attracts households to non-metropolitan counties and much less so to metropolitan counties. The number of heating degree days is always positive for non-metropolitan destinations indicating that the number of heating degree days in winter is an attractive quality. This most likely reflects two different aspects of the allure of non-metropolitan counties: agricultural work and amenity driven migration. For metropolitan-to-metropolitan movement, cold destinations are less desirable. This reflects the long-term outflow of households from the cold winters in northeastern metros of the US and into areas with warmer winters in the western portion of the US. Households moving from non-metropolitan counties and into metropolitan counties also preferred warmer metropolitan counties.

Figure 4.23 on page 178, features the outdoor amenity index coefficients for 1990 through 2015. For non-metropolitan-to-non-metropolitan movement, the generally positive sign of

both the origin and the destination coefficients reflects the proportion of movement between directly adjacent non-metropolitan counties. A non-metropolitan county with a high amenity score is usually grouped with other non-metropolitan counties with high amenity scores. This means that households moving from one high-amenity county are likely to move to an adjacent high-amenity county. For destination coefficients, the type of origin and the type of destination drove the allure amenities. Households moving from non-metropolitan counties were attracted to non-metropolitan destination counties with amenities. High amenity non-metropolitan counties generally promoted an increase in households in the 1990s and dampened incoming flows starting in the 2000s. These periods of attraction correspond to a positive non-metropolitan net migration rate. Outdoor amenities were an attraction in the 1990s but less so after 2000.

4.5 Determinants of metropolitan and non-metropolitan movement

The analysis undertaken in this chapter was made possible by my estimation of the complete set of origins and destinations for 26-years of county-to-county household migration data. By merging the enhanced county-to-county migration data to a consistent set of metropolitan boundaries, the 1990 metropolitan boundaries, I identified four different types of movement: households migrating from non-metropolitan counties to other non-metropolitan counties, households migrating from non-metropolitan counties to metropolitan counties, households migrating from metropolitan counties to non-metropolitan counties, and households migrating from metropolitan counties to other metropolitan counties. I fit over 100 regression models to explain four types of household movement as a function of the age structure, the economic conditions, and the natural amenities in the origin and destination. I make three contributions in this chapter. First, I investigate the spatiality of household movement. The spatial resolution of the county-to-county household migration data enable the identification and the examination of the determinants of four different types of household movement. These determinants include variables pertaining to the spatial structure of counties, the age structure, economic indicators, and outdoor amenity indicators in the origin and destination.

Second, the temporal resolution of the county-to-county migration data enable a comparison of the determinants across a 26-year period, 1990 through 2015. I found that young people ages 20-29 are always the most consistently propulsive and attractive components of a population. Children and teenagers ages 10 through 19 have a dampening effect on both outgoing and incoming flows. Other age groups vary in their propulsive and attractive potential over time. For example, the age 60-69 age group always attracted households to non-metropolitan destination counties during the 1990 through 2015 period. The attractive potential peaked in 1999 and decreased until 2002 when it gradually increased and then plateaued. Other age groups across the four movement types featured a sharp inflection point around 2000.

The economic indicators are mixed. As to be expected, the unemployment rate in the destination acts as dampening force, across all movement types. Over time, the effect of annual average pay in metropolitan counties attracted households from other metropolitan counties while simultaneously dampened flows from non-metropolitan counties. Households were attracted to more affordable non-metropolitan counties and less attracted to less-affordable metropolitan counties. The amenity variables showed the effect of short distance moves the moves. The total inches of precipitation had a small effect on dampening outgoing flows while the total number of heating degree days had a motivating effect on both outgoing and incoming flows and the effect of the outdoor amenity index was mixed. Because the largest share of movement is to directly adjacent counties (approximately 30 miles), across the four movement types, the variation in amenities is going to be slight. Post 2000, households are drawn to metropolitan counties with warmer weather and non-metropolitan counties with cooler weather.

The third contribution I make with this chapter is reflected in the simultaneous opportunity and challenge of using yearly internal migration data. The enhanced county-to-county household migration data feature phenomenal spatial and temporal resolution which present an opportunity in understanding yearly measures of internal migration at the county scale but also present a challenge in finding publicly available datasets with commensurate spatial and temporal resolution. In this regard, I have three recommendations for future study. The

first is to take advantage of more fine-grained metropolitan and non-metropolitan codes such as those provided by rural-urban codes defined by the USDA's Economic Research Service (USDA Economic Research Service 2020). These codes classify counties into nine categories and these categories would help unpack some of the differences in the determinants. The second recommendation is to incorporate additional measures of affordability, amenity, and dis-amenity. The current measure of affordability is only for three distinct years. Partridge et al. (2012) find that housing prices have not fully captured the value of outdoor amenities. If outdoor amenities were captured by housing prices, migration, as the authors claim, would reach a spatial equilibrium. However, Partridge et al. (2012) only investigate positive amenities, pleasant outdoor weather and land forms providing opportunities for outdoor recreation, as do I. Housing prices could also reflect the concentration of environmental disamenity and hazards. Recent work has shown how disamenity and hazards influence migration (Cebula and Alexander 2006; Fussell et al. 2016; Shumway et al. 2014). Future research using the enhanced county-to-county migration data could incorporate measures of environment disamenity and hazards. This is especially timely given climate change and sea level rise. Another reason to incorporate additional measures of affordability is to disentangle the juxtaposition of shorter distance moves and house prices. Presumably, a shorter distance move, approximately 30 miles, still enables a migrant to leverage location specific capital (such as business ties and community networks). Location specific capital is specific to each household and unpacking that would require the use of individual level microdata. The third recommendation is to examine the inflection point in the age categories as seen in 2000. This too would require the use of microdata.

Chapter 5

THE SPATIALITY OF INTERNAL HOUSEHOLD MIGRATION, 1990-2015

Chapters two, three, and four in this dissertation examined the spatiality of internal household migration during the 1990 through 2015 period from three perspectives: the data driving internal migration analysis, the destination preferences of migrating households, and the effect of place-based characteristics on movement between, to, and from metropolitan and non-metropolitan counties. In doing so, I have illustrated how to enhance publicly available migration data, the importance of considering within-state migration in addition to interstate migration, and how place-based characteristics influence household movement.

Chapter two of this dissertation describes the techniques I used to harmonize, reconcile, and enhance a publicly available dataset: the Internal Revenue Service's county-to-county household migration data for years 1990 through 2015. As initially acquired, the county-to-county household migration data are in different file formats that feature an information schema that changes over time. The different file formats and inconsistent schema impede analysis. Harmonizing these data involved the development and instantiation of a database schema accommodating and regularizing the year-over-year differences to produce a dataset with a consistent set of coded migration records for each year in the 1990 through 2015 period. I created a dataset where a single variable indicates if a migration record is within-state or out state and migration to a focal county (or all counties) can be obtained for a single year, a proportion of years, or all years in the 1990 through 2015 period. The reconciliation process added missing county-to-county records so that the sum of all incoming flows and incoming records matches the sum of outgoing flows and outgoing records, respectively. This quality control step balanced flows and records for households and people.

Like data from most federal government agencies, the IRS county-to-county migration data feature measures to protect privacy and confidentiality. In practice, this means that household flows between counties below a certain threshold are suppressed. For data coming from years 1990 through 2012, flows less than 10 households are suppressed and flows less than 20 are suppressed in years 2013 through 2015. The suppressed counts of households migrating into and migrating out of a focal county are included in an aggregate remainder within-state sum or an aggregate remainder interstate sum. In any given year in the study period, approximately 50-percent of internal migration features an origin and a destination within the same state and 25-percent of internal migration features an origin and a destination in different states. Five percent of internal migration is aggregated to the within-state remainder category and 20-percent of internal migration is aggregated to the interstate migration category. In other words, seventy-five percent of the as-reported county-to-county migration data features a fully reported origin and a fully reported destination while the remaining 25-percent feature only a known origin or a known destination, but not both. Figure 2.6 on page 36 features the counts and rates of internal household migration by reported and aggregate remainder categories. The section titled 2.4.2, beginning on page 36, demonstrates the effect of not-including the suppressed county-to-county household flows on the distance decay parameter. The exclusion of these suppressed flows - the omission of records - makes the distance decay parameter more negative giving the impression that interaction decreases more quickly with distance. I developed an algorithm to correct for the potential biases resulting from the omission of suppressed records.

The algorithm I developed features a combination of spatial interaction modelling and linear optimization programming to estimate the origins and destinations of the suppressed household flows between counties. Through the application of the algorithm I developed, I tripled the number of complete within-state household migration records and increased the number of complete interstate household migration records by 13-fold. This means that an additional 1.6 million households, on average, per year, can be included in subsequent analysis. Having more complete cases enables a more accurate analysis of the spatiality of

internal migration for specific origins and destinations as well as the entirety of the migration system. Chapters three and four make use of these enhanced data.

In chapter three, I investigated the destination preferences of migrating households during the 1990 through 2015 period. In particular, the preference for more populous counties, accessible destinations, shorter distance moves, and within-state movement. The determination of these preferences was accomplished through the estimation of approximately 77K production-constrained, origin-specific spatial interaction models. All four dimensions of destination preference show distinct regional trends. Households in counties in the western portion of the US prefer relatively less populated destinations while households in counties in the south and in the central plains prefer relatively more populated destinations. The accessibility of a county is measured as the total number of miles households from all other points of origin would travel if all households were to move a focal county. Counties with greater values of this measure are more accessible to a population than counties with lesser values of this measure. More centrally located counties are more accessible (households would need to travel shorter distances) as are less populous counties (more households could move to a less populated county than vice versa). The distinct and consistent east-west divide of the preference for accessible counties reflects the spatial configuration of counties and the nature of mobility in the late 20th and early 21st century. On average, counties west of the Mississippi River are 2.5 times larger in area than counties east of the Mississippi River which means that, in general, households east of the Mississippi River travel shorter distances when migrating from one county to another. The shorter distances between counties means that counties in the East are more accessible. If a county has a high accessibility measure, there are more destinations with similar or lower accessibility measures than destinations with a higher accessibility measure. The opposite is true with low accessibility measures. Given that a large proportion of household moves are to nearby counties, households in the east demonstrated a preference for relatively less-accessible counties.

The final two destination preferences, geographic distance and within-state movement, warrant joint discussion. In any given year during the study period, approximately 55-percent

of household moves start and end in the same state of origin suggesting that households have a slightly greater preference for within-state destinations and therefore shorter distance moves. Visualizing the results of the spatial interaction models shows distinct patterns. Households in counties in eastern states prefer shorter distance moves. Households in counties in western states and counties in New England and Florida prefer longer distance moves. Similar trends are reflected in the preference for within-state destinations. Households in counties east of the Mississippi River show a strong preference for within-state destinations. That preference, however, decreases in central Ohio, central North Carolina, Florida, and upstate New York. Households in counties in the plains exhibit a strong preference for within-state destinations while households in counties in east Texas, California, and portions of Washington feature less of a relative preference for in-state destinations. Combined, these four coefficients illustrate the outcome of movement between centers of large population, usually metropolitan counties. And to some extent, gravitational theories of human movement would predict some of these patterns - a more populated country is going to exchange more households with other more populated counties. Along those lines, in chapter four I investigated movement between metropolitan and non-metropolitan counties via the age structure, economic indicators, and outdoor amenity indicators in the origin and destination counties.

Chapter four features the results of fitting four models for each year in the 26-year study period. These models explain county-to-county household movement as a function of the place-based characteristics of the origin and the destination. I modeled four different types of household movement: non-metropolitan-to-non-metropolitan, non-metropolitan-to-metropolitan, metropolitan-to-non-metropolitan, and metropolitan-to-metropolitan. Investigating the determinants of four different movement types is made possible by using the enhanced county-to-county household migration data because the enhanced county-to-county household migration feature a greater number of records than the initially available records and therefore enable a more accurate and robust analysis of the households flows by movement type.

The results of modeling the determinants of household movement show how the propulsive

and attractive potentials of each age group and other place-based characteristics change over time and how those potentials vary by movement type. Children and teenagers ages 10 through 19, across movement types showed a consistent ability to dampen outgoing flows of households and incoming flows of households. Young adults ages 20-29, consistently promoted outgoing and incoming flows of households across the four movement types. The other seven age groups feature periods of increasing propulsive and attractive potential and periods of decreasing propulsive and attractive potential. Often, the changes in propulsive and attractive potential reached an inflection point around 2000, across the four movement types.

I conclude by highlighting my contributions and recommendations for future study. In chapter two I described the process I used to harmonize, reconcile, and enhance publicly available data and I make three contributions in this chapter. The first is the development of a framework to harmonize and reconcile publicly available county-to-county household migration data ensuring data quality and consistency. This extract-transform-load (ETL) framework could be applied to other migration datasets or public use datasets. The second contribution I make is the integration of spatial interaction modelling techniques and linear optimization programming techniques to enhance publicly available data. Where appropriate, these techniques could be used on other publicly available data to complete cases. For example, estimating intra-county flows at the tract scale. The American Community Survey details the number of people who moved into each census tract from other parts of the county. Destination totals are known but the origins are not. The technique described in chapter two could be modified to apportion counts of people to the tracts of origin. This in turn could be used to study displacement of vulnerable populations. The third contribution I make in chapter two is the demonstration of the harmonization, reconciling, and enhancement technique for 26-years' worth of county-to-county household migration. In doing so, I have shown how researchers can take advantage of a resource with tremendous temporal and spatial resolution. While this version of the enhanced data features only flows of households, future versions could feature flows of people. Distributing flows of aggregate income

is possible but more difficult on account of some households having a negative income.

I have several recommendations to improve the algorithm I developed to estimate the origins and destinations of the suppressed county-to-county flows could made. First, incorporating techniques for working with censored distributions could be used to provide better estimates of the frequency of flows of specific size. Second, I used a linear optimization programming technique to distribute the counts of households in the aggregate remainder categories and respect the constraints specific to each county. This resulted in a deterministic value. By estimating the origins and destinations, I was completing a matrix. It could be useful to compare the results of other matrix completion techniques (see Davenport and Romberg (2016) for a survey) to see which county-to-county pairs are assigned a value and others are not given various imputation frameworks. I used a linear optimization programming framework because I wanted to respect the maximum county-to-county flow size and ensure that sums of incoming and outgoing flows matched. The framework I developed used mostly open-source software on a 10-year-old laptop. I used Python and R to harmonize and reconcile the data and SQLite to store the harmonized records. Gurobi, the only proprietary software in the tech stack, was used for the linear optimization programming. I was able to take advantage of Gurobi's free academic licensing, however. Future versions of this data will be hosted online and made available to other researchers.

The second empirical chapter featured an analysis of the spatiality of household preferences over time using the enhanced county-to-county household migration data. I used production constrained, origin specific models to investigate the spatiality of household preferences. By showing the preference for within-state destinations, I showed how state borders matter for internal migration studies. It is important to consider both interstate and intrastate flows when examining internal migration. The third empirical chapter used the enhanced county-to-county household migration data to examine how age structure and other place-based characteristics influence household movement. Future models could also incorporate environmental hazards to understand how climate change is influencing migration. For example, Crowell et al. (2010) estimate the number of people living in 100 year

coastal flood plains in the US at about three-percent of the population. With sea level rise, will more people be moving inland? In addition, microdata could be used to further investigate the trends seen in both chapters.

5.0.1 Epilogue: Migration in the time of COVID-19

I wrote most of this dissertation during the COVID-19 pandemic of 2020 and 2021. A time of national quarantine and lock downs. A good proportion of the population in the workforce was instructed to work from home, if able to (Thompson 2020). The pandemic accelerated the rate at which people work from home. Indeed, Cooke (2013) found the rise of communication technologies as a reason for the declining migration rate. With the ability to work from anywhere, people are less tied to specific places as commutes have vanished. Suburban homes are more enticing (Bogost 2020) and amenity-rich small towns are experiencing dramatic increases in population on account of remote work (Smith 2020). For those that can afford it and have the jobs that permit remote work, will people leave metropolitan counties in droves, leaving behind a less affluent and more rooted population? Will the increase in remote work cause non-metropolitan populations to increase? The diffusion of metropolitan residents has political ramifications. For example, Robinson and Noriega (2010) describe how the gains made by the democratic party in the Rocky Mountain West region were driven by voter migration.

When considering the preference for within-state movement (and shorter distance moves), and given Tobler's (1970) first law of geography - near things are more similar than distant things - one might conclude that shorter distance moves translate to encountering populations, spaces, and places more similar to the ones migrants left than not. And moves between metropolitan areas, even distant metropolitan areas, are moves between similar spaces. As Clifford Stoll (2005, p. 308) remarked, reflecting upon his cross country move, "Cambridge, Massachusetts, might be across the country, but culturally, it is just around the corner from Berkley." As households continue to move, will they continue to move to similar spaces or into more different spaces. Is the spatial configuration of internal movement promoting con-

tact with similarity or difference? The IRS has published county-to-county migration data for years 2016, 2017, 2018, and 2019. The 2020 data should be available in the next few years. I look forward to investigating these recent releases of data.

BIBLIOGRAPHY

1992 TIGER/Line® Files [Machine-Readable Data Files]. 1992. ItemType: dataset, Washington DC. https://www2.census.gov/geo/tiger/TIGER1992/.

1999 TIGER/Line® Files [Machine-Readable Data Files]. 1999. ItemType: dataset, Washington DC. https://www2.census.gov/geo/tiger/TIGER1999/.

2000 TIGER/Line® Files [Machine-Readable Data Files]. 2000. Washington DC. https://www2.census.gov/geo/tiger/tiger2k/.

2002 TIGER/Line® Files [Machine-Readable Data Files]. 2002. Washington DC. https://www2.census.gov/geo/tiger/TIGER2002/.

2003 TIGER/Line® Files [Machine-Readable Data Files]. 2003. Washington DC. https://www2.census.gov/geo/tiger/TIGER2003/.

2005 TIGER/Line® Files [Machine-Readable Data Files]. 2005. Washington DC. https://www2.census.gov/geo/tiger/TIGER2005/.

2006 TIGER/Line® Files [Machine-Readable Data Files]. 2006. Washington DC. https://www2.census.gov/geo/tiger/TIGER2006/.

2007 TIGER/Line® Files [Machine-Readable Data Files]. 2007. Washington DC. https://www2.census.gov/geo/tiger/TIGER2007/.

2008 TIGER/Line® Shapefiles [Machine-Readable Data Files]. 2008. Washington DC. https://www2.census.gov/geo/tiger/TIGER2008/.

2009 TIGER/Line® Shapefiles [Machine-Readable Data Files]. 2009. Washington DC. https://www2.census.gov/geo/tiger/TIGER2009/.

2010 TIGER/Line® Shapefiles [Machine-Readable Data Files]. 2010. Washington DC. https://www2.census.gov/geo/tiger/TIGER2010/.

2011 TIGER/Line® Shapefiles [Machine-Readable Data Files]. 2011. Washington DC. https://www2.census.gov/geo/tiger/TIGER2011/.

2012 TIGER/Line® Shapefiles [Machine-Readable Data Files]. 2012. Washington DC. https://www2.census.gov/geo/tiger/TIGER2012/.

2013 TIGER/Line® Shapefiles [Machine-Readable Data Files]. 2013. Washington DC. https://www2.census.gov/geo/tiger/TIGER2013/.

2014 TIGER/Line® Shapefiles [Machine-Readable Data Files]. 2014. Washington DC. https://www2.census.gov/geo/tiger/TIGER2014/.

2015 TIGER/Line® Shapefiles [Machine-Readable Data Files]. 2015. Washington DC. https://www2.census.gov/geo/tiger/TIGER2015/.

Adams, John S., Barbara J. VanDrasek, and Eric G. Phillips. 1999. "METROPOLITAN AREA DEFINITION IN THE UNITED STATES". *Urban Geography* 20, no. 8 (): 695–726. ISSN: 0272-3638, 1938-2847, visited on 04/21/2021. doi:10.2747/0272-3638.20.8.695. http://www.tandfonline.com/doi/abs/10.2747/0272-3638.20.8.695.

Agresta, A. 1985. "The migration turnaround: end of a phenomenon?" *Population Today* 13, no. 1 (): 6–7. ISSN: 0749-2448.

Akwawua, Siaw, and James A Pooler. 2001. "The development of an intervening opportunities model with spatial dominance effects". *Journal of Geographical Systems* 3 (1): 69–86. ISSN: 1435-5930.

Anderson, Theodore R. 1955. "Intermetropolitan Migration: A Comparison of the Hypotheses of Zipf and Stouffer". Publisher: [American Sociological Association, Sage Publications, Inc.] *American Sociological Review* 20 (3): 287–291. ISSN: 00031224, visited on 06/30/2020. doi:10.2307/2087387. www.jstor.org/stable/2087387.

— . 1956. "Intermetropolitan Migration: A Correlation Analysis". Publisher: The University of Chicago Press, *American Journal of Sociology* 61, no. 5 (): 459–462. ISSN: 0002-9602, visited on 06/30/2020. doi:10.1086/221805. https://doi.org/10.1086/221805.

Atlas Van Lines. 2020. *Atlas Van Lines - Migration Patterns — Atlas Van Lines*. Visited on 07/04/2021. https://www.atlasvanlines.com/migration-patterns.

Babyak, Michael A. 2004. "What You See May Not Be What You Get: A Brief, Nontechnical Introduction to Overfitting in Regression-Type Models". *Psychosomatic Medicine*: 11.

Ballard, Patricia L., and Glenn V. Fuguitt. 1985. "The changing small town settlement structure in the United States, 1900-1980". Publisher: Rural Sociological Society, etc. *Rural Sociology* 50 (1): 99. ISSN: 0036-0112.

Bell, Martin, et al. 2015. "Internal migration and development: comparing migration intensities around the world". *Population and Development Review* 41 (1): 33–58. ISSN: 0098-7921.

Bernard, Aude, Bell, Martin, and Charles-Edwards, Elin. 2014. "Life-Course Transitions and the Age Profile of Internal Migration". *Population and Development Review* 40 (2): 213–239. http://www.jstor.org/stable/24027921.

Berry, Brian JL. 1976. "The counterurbanization process: urban America since 1970". Publisher: Sage, *Urbanization & counterurbanization*: 17–30.

Beyers, William B, and David P Lindahl. 1996. "Lone eagles and high fliers in rural producer services". *Rural America/Rural Development Perspectives* 11 (2221-2019-2629): 2–10.

Bilsborrow, Richard E., and John S. Akin. 1982. "Data Availability versus Data Needs for Analyzing the Determinants and Consequences of Internal Migration: An Evaluation of U.S. Survey Data." *Review of Public Data Use* 10, no. 4 (): 261. ISSN: 00922846. http://offcampus.lib.washington.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=6639279&site=ehost-live.

Bogost, Ian. 2020. *Revenge of the Suburbs*. Section: Technology. Visited on 06/16/2021. https://www.theatlantic.com/technology/archive/2020/06/pandemic-suburbs-are-best/613300/.

Brewer, Cynthia A. 2005. *Designing better maps : a guide for GIS users*. 1st ed. Redlands, Calif.: Redlands, Calif. : ESRI Press.

Bureau of Labor Statistics. 2020. *Local Area Unemployment Statistics Estimation Methodology*. Visited on 06/28/2021. https://www.bls.gov/lau/laumthd.htm.

— . 2021. *Quarterly Census of Employment and Wages*. https://www.bls.gov/opub/hom/cew/pdf/cew.pdf.

Cadieux, Kirsten Valentine, and Patrick T Hurley. 2011. "Amenity migration, exurbia, and emerging rural landscapes: Global natural amenity as place and as process". Publisher: Springer, *GeoJournal* 76 (4): 297–302. ISSN: 0343-2521.

Carrothers, Gerald A. P. 1956. "An Historical Review of the Gravity and Potential Concepts of Human Interaction". Publisher: Routledge, *Journal of the American Institute of Planners* 22, no. 2 (): 94–102. ISSN: 0002-8991. doi:10.1080/01944365608979229. https://doi.org/10.1080/01944365608979229.

Castro, Luis J, and Andrei Rogers. 1984. "What the age composition of migrants can tell us". Publisher: RR-84-003. Reprinted from Population Bulletin of the United Nations 1983, *Population Bulletin of the United Nations 1983*.

Cebula, Richard J, and Gigi M Alexander. 2006. "Determinants of net interstate migration, 2000-2004". *Journal of Regional Analysis and Policy* 36 (1100-2016-89796).

Clark, G L, and K P Ballard. 1980. "Modeling out-Migration from Depressed Regions: The Significance of Origin and Destination Characteristics". *Environment and Planning A: Economy and Space* 12, no. 7 (): 799–812. ISSN: 0308-518X, 1472-3409, visited on 07/09/2021. doi:10.1068/a120799. http://journals.sagepub.com/doi/10.1068/a120799.

Clark, William A V, and Suzanne Davies Withers. 2009. "Fertility, mobility and labour-force participation: a study of synchronicity". *Population, Space and Place* 15 (4): 305–321. ISSN: 1544-8452. doi:10.1002/psp.555. http://dx.doi.org/10.1002/psp.555.

Cooke, Thomas J. 2013. "Internal Migration in Decline". *The Professional Geographer* 65, no. 4 (): 664–675. ISSN: 0033-0124, 1467-9272, visited on 05/13/2016. doi:10.1080/00330124.2012.724343. http://www.tandfonline.com/doi/abs/10.1080/00330124.2012.724343.

Cooke, Thomas J., Richard Wright, and Mark Ellis. 2018. "A Prospective on Zelinsky's Hypothesis of the Mobility Transition". *Geographical Review* 108, no. 4 (): 503–522. ISSN: 0016-7428, 1931-0846, visited on 07/08/2021. doi:10.1111/gere.12310. https://www.tandfonline.com/doi/full/10.1111/gere.12310.

Cromartie, John, and Shawn Bucholtz. 2008. "Defining the" rural" in rural America". *Amber Waves* 6 (3): 29–34.

Crowell, Mark, et al. 2010. "An Estimate of the U.S. Population Living in 100-Year Coastal Flood Hazard Areas". *Journal of Coastal Research* 262 (): 201–211. ISSN: 0749-0208, 1551-5036, visited on 02/28/2019. doi:10.2112/JCOASTRES-D-09-00076.1. http://www.bioone.org/doi/abs/10.2112/JCOASTRES-D-09-00076.1.

Curry, Leslie. 1972. "A spatial analysis of gravity flows". Publisher: Taylor & Francis, *Regional Studies* 6 (2): 131–147. ISSN: 0034-3404.

Curry, Leslie, Daniel A Griffith, and Eric S Sheppard. 1975. "Those gravity parameters again". Publisher: Taylor & Francis, *Regional Studies* 9 (3): 289–296. ISSN: 0034-3404.

Davenport, Mark A., and Justin Romberg. 2016. "An overview of low-rank matrix recovery from incomplete observations". ArXiv: 1601.06422, *IEEE Journal of Selected Topics in Signal Processing* 10, no. 4 (): 608–622. ISSN: 1932-4553, 1941-0484, visited on 03/21/2020. doi:10.1109/JSTSP.2016.2539100. http://arxiv.org/abs/1601.06422.

Dennett, Adam. 2012. "Estimating flows between geographical locations: 'get me started in' spatial interaction modelling": 25.

DeWaard, Jack, Janna E. Johnson, and Stephan D. Whitaker. 2018. *Internal Migration in the United States: A Comparative Assessment of the Utility of the Consumer Credit Panel*. Working paper (Federal Reserve Bank of Cleveland) WP 18-04. Federal Reserve Bank of Cleveland. Visited on 01/19/2020. doi:10.26509/frbc-wp-201804. https://www.clevelandfed.org/en/newsroom-and-events/publications/working-papers/2018-working-papers/wp-1804-internal-migration-in-the-united-states.aspx.

DeWaard, Jack, et al. 2020a. "Changing spatial interconnectivity during the "Great American Migration Slowdown": A decomposition of intercounty migration rates, 1990-2010". *Population, Space and Place* 26, no. 1 (). ISSN: 1544-8444, 1544-8452, visited on 06/24/2021. doi:10.1002/psp.2274. https://onlinelibrary.wiley.com/doi/abs/10.1002/psp.2274.

DeWaard, Jack, et al. 2020b. "User Beware: Concerning Findings from Recent U.S. Internal Revenue Service Migration Data". Medium: PDF Publisher: Minneapolis, MN: Minnesota Population Center, *Minnesota Population Center Working Paper Series*. Visited on 08/16/2020. doi:10.18128/MPC2020-02. https://assets.ipums.org/_files/mpc/wp2020-02.pdf.

Dillman, Don A. 1979. "Residential Preferences, Quality of Life, and the Population Turnaround". *American Journal of Agricultural Economics* 61, no. 5 (): 960–966. ISSN: 0002-9092, 1467-8276, visited on 06/22/2021. doi:10.2307/3180356. https://onlinelibrary.wiley.com/doi/abs/10.2307/3180356.

Dorigo, Guido, and Waldo Tobler. 1983. "Push-Pull Migration Laws". *Annals of the Association of American Geographers* 73 (1): 1–17. ISSN: 0004-5608.

Duncombe, William, Mark Robbins, and Douglas A Wolf. 2001. "Retire to where? A discrete choice model of residential location". *International Journal of Population Geography* 7 (4): 281–293. ISSN: 1099-1220. doi:10.1002/ijpg.227. http://dx.doi.org/10.1002/ijpg.227.

ESRI Shapefile Technical Description. 1998.

Farley, John E. 2007. "Metropolitan Statistical Area". In *The Blackwell Encyclopedia of Sociology*, ed. by George Ritzer, wbeosm095. Oxford, UK: John Wiley & Sons, Ltd. ISBN: 978-1-4051-2433-1, visited on 04/19/2021. doi:10.1002/9781405165518.wbeosm095. http://doi.wiley.com/10.1002/9781405165518.wbeosm095.

Flood, Sarah, et al. 2018. *Integrated Public Use Microdata Series, Current Population Survey: Version 6.0*. Type: dataset. Visited on 01/18/2019. doi:10.18128/D030.V6.0. http://cps.ipums.org.

Flowerdew, Robin. 1982. "Fitting the Lognormal Gravity Model to Heteroscedastic Data". *Geographical Analysis* 14 (3): 263–267. ISSN: 1538-4632. doi:10.1111/j.1538-4632.1982.tb00075.x. http:https://doi.org/10.1111/j.1538-4632.1982.tb00075.x.

Flowerdew, Robin, and Murray Aitkin. 1982. "A METHOD OF FITTING THE GRAVITY MODEL BASED ON THE POISSON DISTRIBUTION." *Journal of Regional Science* 22, no. 2 (): 191. ISSN: 00224146. http://offcampus.lib.washington.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=5749994&site=ehost-live.

Foster, Thomas B., Mark J. Ellis, and Lee Fiorio. 2018. "Foreign-born and native-born migration in the U.S.: evidence from linked IRS administrative and census survey records". *Journal of Population Research* 35, no. 4 (): 467–498. ISSN: 1835-9469. doi:10.1007/s12546-018-9215-x. https://doi.org/10.1007/s12546-018-9215-x.

Fotheringham, A. Stewart. 1981. "Spatial Structure and Distance-Decay Parameters". Publisher: [Association of American Geographers, Taylor & Francis, Ltd.] *Annals of the Association of American Geographers* 71 (3): 425–436. ISSN: 00045608, 14678306, visited on 05/18/2020. www.jstor.org/stable/2562901.

Fotheringham, A Stewart. 1983. "A new set of spatial-interaction models: the theory of competing destinations". *Environment and Planning A: Economy and Space* 15 (1): 15–36. ISSN: 0308-518X.

— . 1986. "Further discussion on distance-deterrence parameters and the competing destinations model". *Environment and planning A* 18 (4): 553–556. ISSN: 0308-518X.

Fotheringham, A Stewart, and Michael J Webber. 1980. "Spatial structure and the parameters of spatial interaction models". Publisher: Blackwell Publishing Ltd Oxford, UK, *Geographical Analysis* 12 (1): 33–46. ISSN: 0016-7363.

Foulkes, Matthew, and K Bruce Newbold. 2008. "Using alternative data sources to study rural migration: examples from Illinois". *Population, Space and Place* 14 (3): 177–188. ISSN: 1544-8452. doi:10.1002/psp.482. http://dx.doi.org/10.1002/psp.482.

Frey, William H. 1993. "The new urban revival in the United States". *Urban Studies* 30 (4-5): 741–774. ISSN: 0042-0980.

Fuguitt, Glenn V. 1985. "The Nonmetropolitan Population Turnaround". *Annual Review of Sociology* 11, no. 1 (): 259–280. ISSN: 0360-0572, 1545-2115, visited on 06/22/2021. doi:10.1146/annurev.so.11.080185.001355. http://www.annualreviews.org/doi/10.1146/annurev.so.11.080185.001355.

Fuguitt, Glenn V., and Calvin L. Beale. 1996. "Recent Trends in Nonmetropolitan Migration: Toward a New Turnaround?" *Growth and Change* 27, no. 2 (): 156–174. ISSN: 0017-4815, 1468-2257, visited on 06/22/2021. doi:10.1111/j.1468-2257.1996.tb00901.x. http://doi.wiley.com/10.1111/j.1468-2257.1996.tb00901.x.

Fussell, Elizabeth, et al. 2016. "Weather-Related Hazards and Population Change: A Study of Hurricanes and Tropical Storms in the United States, 1980-2012". *The ANNALS of the American Academy of Political and Social Science* 669, no. 1 (): 146–167. ISSN: 0002-7162, visited on 02/15/2018. doi:10.1177/0002716216682942. https://doi.org/10.1177/0002716216682942.

G. T. M. 1932. "LONG LINES ON THE EARTH". *Empire Survey Review* 1, no. 6 (): 259–263. ISSN: 0267-1034, visited on 06/21/2020. doi:10.1179/sre.1932.1.6.259. http://www.tandfonline.com/doi/full/10.1179/sre.1932.1.6.259.

GDAL/OGR contributors. 2019. *GDAL/OGR Geospatial Data Abstraction Software Library*. https://gdal.org.

Gholamy, Afshin, Vladik Kreinovich, and Olga Kosheleva. 2018. "Why 70/30 or 80/20 Relation Between Training and Testing Sets: A Pedagogical Explanation". University of Texas at El Paso.

Gillies, Sean. 2018. *Shapely*. https://shapely.readthedocs.io/en/latest/.

Glaeser, Edward L, and Jesse M Shapiro. 2003. "Urban Growth in the 1990s: Is City Living Back?" *Journal of Regional Science* 43, no. 1 (): 139–165. ISSN: 0022-4146, 1467-9787, visited on 07/09/2021. doi:10.1111/1467-9787.00293. https://onlinelibrary.wiley.com/doi/10.1111/1467-9787.00293.

Graves, Philip E. 1979. "A life-cycle empirical analysis of migration and climate, by race". *Journal of Urban Economics* 6, no. 2 (): 135–147. ISSN: 00941190, visited on 07/09/2021. doi:10.1016/0094-1190(79)90001-9. https://linkinghub.elsevier.com/retrieve/pii/0094119079900019.

Greenwood, Michael J., and Gary L. Hunt. 1989. "Jobs versus amenities in the analysis of metropolitan migration". *Journal of Urban Economics* 25, no. 1 (): 1–16. ISSN: 00941190, visited on 07/09/2021. doi:10.1016/0094-1190(89)90040-5. https://linkinghub.elsevier.com/retrieve/pii/0094119089900405.

Greenwood, Michael J., et al. 1991. "Migration, Regional Equilibrium, and the Estimation of Compensating Differentials". *The American Economic Review* 81 (5): 1382–1390. ISSN: 00028282. http://www.jstor.org/stable/2006927.

Griffith, Daniel A. 1982. "Geometry and Spatial Interaction". *Annals of the Association of American Geographers* 72, no. 3 (): 332–346. ISSN: 0004-5608, 1467-8306, visited on 09/02/2019. doi:10.1111/j.1467-8306.1982.tb01829.x. http://www.tandfonline.com/doi/abs/10.1111/j.1467-8306.1982.tb01829.x.

Gross, Emily. 2005. "Internal revenue service area-to-area migration data: strengths, limitations, and current uses". *Statistics of Income. SOI Bulletin* 25 (3): 159–160. ISSN: 0730-0743.

— . 2009. *U.S. Population Migration Data: Strengths and Limitations*. Visited on 09/17/2018. https://www.irs.gov/pub/irs-soi/99gross_update.doc.

Hauer, Mathew, and James Byars. 2019. "IRS county-to-county migration data, 1990 - 2010". *Demographic Research* 40 (): 1153–1166. ISSN: 1435-9871, visited on 04/09/2021. doi:10.4054/DemRes.2019.40.40. https://www.demographic-research.org/volumes/vol40/40/.

Henrie, Christopher, and David Plane. 2008. "Exodus from the California Core: Using Demographic Effectiveness and Migration Impact Measures to Examine Population Redistribution Within the Western United States". *POPULATION RESEARCH AND POLICY REVIEW* 27 (1): 43–64. ISSN: 0167-5923. doi:10.1007/s11113-007-9053-6. http://dx.doi.org/10.1007/s11113-007-9053-6.

Herting, Jerald R., David B. Grusky, and Stephen E. Van Rompaey. 1997. "The Social Geography of Interstate Mobility and Persistence". *American Sociological Review* 62, no. 2 (): 267. ISSN: 00031224, visited on 07/09/2021. doi:10.2307/2657304. http://www.jstor.org/stable/2657304?origin=crossref.

Hipp, Dwayne Richard. 2020. *SQLite*. https://www.sqlite.org/index.html.

Internal Revenue Service. 2018. *U.S. Population Migration Data*. https://www.irs.gov/statistics/soi-tax-stats-migration-data.

IPUMS. 2016. *GIS Files*. Text. Visited on 07/25/2019. https://www.nhgis.org/documentation/gis-data.

IRS. 2016. *2013 - 2014 Migration Data Users Guide*. https://www.irs.gov/pub/irs-soi/1314inpublicmigdoc.pdf.

Isserman, Andrew M. 2005. "In the National Interest: Defining Rural and Urban Correctly in Research and Public Policy". *International Regional Science Review* 28, no. 4 (): 465–499. ISSN: 0160-0176, 1552-6925, visited on 04/21/2021. doi:10.1177/0160017605279000. http://journals.sagepub.com/doi/10.1177/0160017605279000.

Isserman, Andrew M., David A. Plane, and David B. McMillen. 1982. "Internal Migration in the United States: An Evaluation of Federal Data." *Review of Public Data Use* 10, no. 4 (): 285. ISSN: 00922846. http://offcampus.lib.washington.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=6639286&site=ehost-live.

Isserman, Andrew M., and James Westervelt. 2006. "1.5 Million Missing Numbers: Overcoming Employment Suppression in County Business Patterns Data". Publisher: SAGE Publications Inc, *International Regional Science Review* 29, no. 3 (): 311–335. ISSN: 0160-0176, visited on 06/04/2021. doi:10.1177/0160017606290359. https://doi.org/10.1177/0160017606290359.

Jenkins, David G., and Pedro F. Quintana-Ascencio. 2020. "A solution to minimum sample size for regressions". Ed. by Gang Han. *PLOS ONE* 15, no. 2 (): e0229345. ISSN: 1932-6203, visited on 06/10/2021. doi:10.1371/journal.pone.0229345. https://dx.plos.org/10.1371/journal.pone.0229345.

Johnson, Kenneth M., and Daniel T. Lichter. 2019. "Rural Depopulation: Growth and Decline Processes over the Past Century". Publisher: John Wiley & Sons, Ltd, *Rural Sociology* 84, no. 1 (): 3–27. ISSN: 0036-0112, visited on 08/07/2020. doi:10.1111/ruso.12266. https://doi.org/10.1111/ruso.12266.

— . 2020. “Metropolitan Reclassification and the Urbanization of Rural America”. *Demography* 57, no. 5 (): 1929–1950. ISSN: 0070-3370, 1533-7790, visited on 04/21/2021. doi:10.1007/s13524-020-00912-5. https://read.dukeupress.edu/demography/article/57/5/1929/168384/Metropolitan-Reclassification-and-the-Urbanization.

Johnson, Kenneth M., et al. 2005. “Temporal and spatial variation in age-specific net migration in the United States”. *Demography* 42, no. 4 (): 791–812. ISSN: 0070-3370, 1533-7790, visited on 06/22/2021. doi:10.1353/dem.2005.0033. https://read.dukeupress.edu/demography/article/42/4/791/170293/Temporal-and-spatial-variation-in-age-specific-net.

Johnston, Ronald J. 1973. “On frictions of distance and regression coefficients”. Publisher: JSTOR, *Area*: 187–191. ISSN: 0004-0894.

Kaplan, Greg, and Sam Schulhofer-Wohl. 2012. “Interstate Migration Has Fallen Less Than You Think: Consequences of Hot Deck Imputation in the Current Population Survey”. *Demography* 49, no. 3 (): 1061–74. ISSN: 00703370. doi:10.1007/s13524-012-0110-3. http://search.proquest.com/docview/1026558364?accountid=14784.

— . 2017. “Understanding the long-run decline in interstate migration”. Publisher: Wiley Online Library, *International Economic Review* 58 (1): 57–94. ISSN: 0020-6598.

Klove, Robert C. 1952. “The Definition of Standard Metropolitan Areas”. Publisher: Routledge, *Economic Geography* 28, no. 2 (): 95–104. ISSN: 0013-0095. doi:10.2307/141019. https://www.tandfonline.com/doi/abs/10.2307/141019.

Kone, Zovanga L, et al. 2018. “Internal borders and migration in India”. Publisher: Oxford University Press, *Journal of Economic Geography* 18 (4): 729–759. ISSN: 1468-2702.

Liaw, Kao-Lee, and William H Frey. 1998. “Destination choices of the 1985-90 young adult immigrants to the United States: Importance of race, educational attainment, and labour market forces”. *International Journal of Population Geography* 4 (1): 49–61. ISSN: 1099-1220. doi:10.1002/(sici)1099-1220(199803)4:1<49::aid-ijpg84>3.0.co;2-

n. http://dx.doi.org/10.1002/(SICI)1099-1220(199803)4:1%3C49::AID-IJPG84%3E3.0.CO;2-N.

— . 2007. "Multivariate explanation of the 1985-1990 and 1995-2000 destination choices of newly arrived immigrants in the United States: the beginning of a new trend?" *Population, Space and Place* 13 (5): 377–399. ISSN: 1544-8452. doi:10.1002/psp.459. http://dx.doi.org/10.1002/psp.459.

Lin, Allen Yilun, Justin Cranshaw, and Scott Counts. 2019. "Forecasting U.S. Domestic Migration Using Internet Search Queries". In *The World Wide Web Conference on - WWW '19*, 1061–1072. San Francisco, CA, USA: ACM Press. ISBN: 978-1-4503-6674-8, visited on 07/04/2021. doi:10.1145/3308558.3313667. http://dl.acm.org/citation.cfm?doid=3308558.3313667.

Little, Jani S, and Andrei Rogers. 2007. "What can the age composition of a population tell us about the age composition of its out-migrants?" *Population, Space and Place* 13 (1): 23–39. ISSN: 1544-8452. doi:10.1002/psp.440. http://dx.doi.org/10.1002/psp.440.

Long, L., and A. Nucci. 1998. "Accounting for two population turnarounds in nonmetropolitan America". *Research in Rural Sociology and Development* 7:47–70. ISSN: 1057-1922.

Long, Larry, and Diana DeAre. 1988. "US Population Redistribution: A Perspective on the Nonmetropolitan Turnaround". *Population and Development Review* 14, no. 3 (): 433. ISSN: 00987921, visited on 06/22/2021. doi:10.2307/1972197. https://www.jstor.org/stable/1972197?origin=crossref.

Lundholm, Emma. 2012. "Returning home? Migration to birthplace among migrants after age 55". *Population, Space and Place* 18 (1): 74–84. ISSN: 1544-8452. doi:10.1002/psp.645. http://dx.doi.org/10.1002/psp.645.

Manson, Steven et al. 2020. *National Historical Geographic Information System: Version 15.0*. Version Number: 15.0 Type: dataset. Visited on 07/02/2021. doi:10.18128/D050.V15.0. https://www.ipums.org/projects/ipums-nhgis/d050.V15.0.

Manson, Gary A., and Richard E. Groop. 2000. "U.S. Intercounty Migration in the 1990s: People and Income Move Down the Urban Hierarchy". *The Professional Geographer* 52, no. 3 (): 493–504. ISSN: 0033-0124, 1467-9272, visited on 04/09/2021. doi:10.1111/0033-0124.00241. http://www.tandfonline.com/doi/abs/10.1111/0033-0124.00241.

McGranahan, David A. 1999. *Natural amenities drive rural population change.*

McGranahan, David A, John Cromartie, and Timothy R Wojan. 2011. *The two faces of rural population loss through outmigration.* Tech. rep.

McKinney, Wes. 2010. "Data Structures for Statistical Computing in Python". In *Proceedings of the 9th Python in Science Conference*, ed. by Stéfan van der Walt and Jarrod Millman, 51–56.

Miller, Edward. 1973. "Is Out-Migration Affected by Economic Conditions?" *Southern Economic Journal* 39, no. 3 (): 396. ISSN: 00384038, visited on 04/08/2021. doi:10.2307/1056406. https://www.jstor.org/stable/1056406?origin=crossref.

Mitchell, Clare J.A. 2004. "Making sense of counterurbanization". *Journal of Rural Studies* 20, no. 1 (): 15–34. ISSN: 07430167, visited on 04/14/2021. doi:10.1016/S0743-0167(03)00031-7. https://linkinghub.elsevier.com/retrieve/pii/S0743016703000317.

Molloy, Raven, Christopher L. Smith, and Abigail Wozniak. 2011. "Internal Migration in the United States". *The Journal of Economic Perspectives* 25 (3): 173–196. ISSN: 08953309. http://www.jstor.org/stable/23049428.

Morrill, Richard, John Cromartie, and Gary Hart. 1999. "METROPOLITAN, URBAN, AND RURAL COMMUTING AREAS: TOWARD A BETTER DEPICTION OF THE UNITED STATES SETTLEMENT SYSTEM". *Urban Geography* 20, no. 8 (): 727–748. ISSN: 0272-3638, 1938-2847, visited on 04/21/2021. doi:10.2747/0272-3638.20.8.727. http://www.tandfonline.com/doi/abs/10.2747/0272-3638.20.8.727.

Nelson, Lise, and Peter B Nelson. 2011. "The global rural: Gentrification and linked migration in the rural USA". *PROGRESS IN HUMAN GEOGRAPHY* 35 (4): 441–459. doi:10.1177/0309132510380487. http://phg.sagepub.com/cgi/content/abstract/35/4/441.

Nelson, Peter B., Ahn Wei Lee, and Lise Nelson. 2009. "Linking baby boomer and Hispanic migration streams into rural America - a multi-scaled approach". *Population, Space and Place* 15, no. 3 (): 277–293. ISSN: 15448444, 15448452, visited on 07/27/2020. doi:10.1002/psp.520. http://doi.wiley.com/10.1002/psp.520.

Newbold, K Bruce. 2010. *Population geography: tools and issues*. Lanham: Rowman & Littlefield Publishers. ISBN: 978-0-7425-5753-6.

Nijkamp, P. 1979. "Gravity and entropy models: the state of the art". *Research-Memorandum*, **numbers** 1977-2.

North American Datum of 1983. 2018. https://www.ngs.noaa.gov/datums/horizontal/north-american-datum-1983.shtml.

North American Van Lines. 2020. *2020 U.S. Moving Migration Patterns Report — North American Van Lines*. Visited on 07/04/2021. https://www.northamerican.com/migration-map.

Office of Management and Budget. 1998. "Alternative Approaches to Defining Metropolitan and Nonmetropolitan Areas". *Federal Register* 63, no. 244 (): 70526–70561. https://www.federalregister.gov/documents/1998/12/21/98-33676/alternative-approaches-to-defining-metropolitan-and-nonmetropolitan-areas.

— . 2000. "Standards for Defining Metropolitan and Micropolitan Statistical Areas". *Federal Register* 65, no. 249 (): 82228–82238. Visited on 04/21/2021. https://www.federalregister.gov/documents/2000/12/27/00-32997/standards-for-defining-metropolitan-and-micropolitan-statistical-areas.

— . 2010. “2010 Standards for Delineating Metropolitan and Micropolitan Statistical Areas; Notice”. *Federal Register* 75, no. 123 (): 8.

— . 2021. *Recommendations from the Metropolitan and Micropolitan Statistical Area Standards Review Committee to the Office of Management and Budget: Changes to the 2010 Standards for Delineating Metropolitan and Micropolitan Statistical Areas*. Visited on 05/29/2021. https://www.regulations.gov/document/OMB-2021-0001-0001.

Olsson, Gunnar. 1970. “Explanation, Prediction, and Meaning Variance: An Assessment of Distance Interaction Models”. Publisher: Routledge, *Economic Geography* 46, no. sup1 (): 223–233. ISSN: 0013-0095. doi:10.2307/143140. https://www.tandfonline.com/doi/abs/10.2307/143140.

Pandit, Kavita. 1997. “Cohort and Period Effects in U.S. Migration: How Demographic and Economic Cycles Influence the Migration Schedule”. *Annals of the Association of American Geographers* 87, no. 3 (): 439–450. ISSN: 0004-5608, 1467-8306, visited on 06/04/2021. doi:10.1111/1467-8306.00062. http://www.tandfonline.com/doi/abs/10.1111/1467-8306.00062.

Partridge, Mark D., et al. 2012. “Dwindling U.S. internal migration: Evidence of spatial equilibrium or structural shifts in local labor markets?” *Regional Science and Urban Economics* 42, **numbers** 1-2 (): 375–388. ISSN: 01660462, visited on 05/30/2018. doi:10.1016/j.regsciurbeco.2011.10.006. http://linkinghub.elsevier.com/retrieve/pii/S0166046211001281.

Pellegrini, Pasquale A., and A. Stewart Fotheringham. 2002. “Modelling spatial choice: a review and synthesis in a migration context.” *Progress in Human Geography* 26, no. 4 (): 487–510. ISSN: 03091325. http://offcampus.lib.washington.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=a9h&AN=7008851&site=ehost-live.

Phillips, Martin. 2009. "Counterurbanisation and rural gentrification: an exploration of the terms". *Population, Space and Place*: n/a–n/a. ISSN: 15448444, 15448452, visited on 04/14/2021. doi:10.1002/psp.570. http://doi.wiley.com/10.1002/psp.570.

Pierce, Kevin. 2015. "SOI Migration Data: A New Approach": 4.

— . 2020. *Note re the 2014-15 IRS Migration Data*. http://mcdc.missouri.edu/data/irsmig/Note_re1415data.html.

Pipa, Anthony F., and Natalie Geismar. 2021. *The new 'rural'? The implications of OMB's proposal to redefine nonmetro America*. Visited on 04/21/2021. https://www.brookings.edu/research/the-new-rural-the-implications-of-ombs-proposal-to-redefine-nonmetro-america/.

Plane, D A, C J Henrie, and M J Perry. 2005. "Migration up and down the urban hierarchy and across the life course". *Proceedings of the National Academy of Sciences of the United States of America* 102 (43): 15313–15318. http://www.pnas.org/content/102/43/15313.abstract.

Plane, David A. 1984. "Migration Space: Doubly Constrained Gravity Model Mapping of Relative Interstate Separation". *Annals of the Association of American Geographers* 74, no. 2 (): 244–256. ISSN: 0004-5608, 1467-8306, visited on 09/02/2019. doi:10.1111/j.1467-8306.1984.tb01451.x. http://www.tandfonline.com/doi/abs/10.1111/j.1467-8306.1984.tb01451.x.

— . 1993. "Demographic Influences on Migration". Publisher: Routledge, *Regional Studies* 27, no. 4 (): 375–383. ISSN: 0034-3404. doi:10.1080/00343409312331347635. https://doi.org/10.1080/00343409312331347635.

— . 1999. "Migration Drift". *The Professional Geographer* 51, no. 1 (): 1–11. ISSN: 0033-0124. doi:10.1111/0033-0124.00140. https://doi.org/10.1111/0033-0124.00140.

Plane, David A., and Frank Heins. 2003. "Age articulation of U.S. inter-metropolitan migration flows". *The Annals of Regional Science* 37, no. 1 (): 107–130. ISSN: 0570-1864, 1432-0592, visited on 07/09/2021. doi:10.1007/s001680200114. http://link.springer.com/10.1007/s001680200114.

Plane, David A., and Gordon F. Mulligan. 1997. "Measuring spatial focusing in a migration system". *Demography* 34, no. 2 (): 251–262. ISSN: 1533-7790. doi:10.2307/2061703. https://doi.org/10.2307/2061703.

Plane, David A., and Peter A. Rogerson. 1991. "TRACKING THE BABY BOOM, THE BABY BUST, AND THE ECHO GENERATIONS: HOW AGE COMPOSITION REGULATES US MIGRATION". *The Professional Geographer* 43, no. 4 (): 416–430. ISSN: 0033-0124, 1467-9272, visited on 07/09/2021. doi:10.1111/j.0033-0124.1991.00416.x. http://www.tandfonline.com/doi/abs/10.1111/j.0033-0124.1991.00416.x.

Pooler, Jim. 1994. "An extended family of spatial interaction models". *Progress in Human Geography* 18, no. 1 (): 17–39. ISSN: 0309-1325, visited on 02/01/2018. doi:10.1177/030913259401800102. https://doi.org/10.1177/030913259401800102.

Population Division. 2000. *(CO-99-9) Population Estimates for Counties by Age and Sex: Annual Time Series July 1, 1990 to July 1, 1999*. Visited on 01/12/2021. https://www2.census.gov/programs-surveys/popest/tables/1990-2000/counties/asrh/.

— . 2016. *(CO-00-9) Population Estimates for Counties by Age and Sex: Annual Time Series July 1, 2000 to July 1, 2010*. Visited on 01/12/2021. https://www2.census.gov/programs-surveys/popest/datasets/2000-2010/intercensal/county/co-est00int-agesex-5yr.csv.

— . 2020. *CC-EST2019-AGESEX-[ST-FIPS]: Annual County and Puerto Rico Municipio Resident Population Estimates by Selected Age Groups and Sex: April 1, 2010 to July 1, 2019*. Visited on 01/12/2021. https://www2.census.gov/programs-surveys/popest/datasets/2000-2010/intercensal/county/co-est00int-agesex-5yr.csv.

Rappaport, Jordan. 2007. "Moving to nice weather". *Regional Science and Urban Economics* 37, no. 3 (): 375–398. ISSN: 01660462, visited on 07/08/2021. doi:10.1016/j.regsciurbeco.2006.11.004. https://linkinghub.elsevier.com/retrieve/pii/S0166046206001001.

Ratcliffe, Michael R. 2002. "Creating metropolitan and micropolitan statistical areas". In *Measuring Rural Diversity*, 3:1–13. Chicago.

Ravenstein, E G. 1885. "The Laws of Migration". *Journal of the Statistical Society of London* 48 (2): 167–235. ISSN: 09595341. http://www.jstor.org/stable/2979181.

— . 1889. "The Laws of Migration". *Journal of the Royal Statistical Society* 52 (2): 241–305. ISSN: 0952-8385.

Reichert, Christiane von, John B. Cromartie, and Ryan O. Arthun. 2014. "Reasons for Returning and Not Returning to Rural U.S. Communities". *The Professional Geographer* 66, no. 1 (): 58–72. ISSN: 0033-0124, 1467-9272, visited on 04/14/2021. doi:10.1080/00330124.2012.725373. https://www.tandfonline.com/doi/full/10.1080/00330124.2012.725373.

Robinson, Tony, and Stephen Noriega. 2010. "Voter migration as a source of electoral change in the Rocky Mountain West". *Political Geography* 29 (1): 28–39. ISSN: 0962-6298. doi:10.1016/j.polgeo.2009.12.012. http://www.sciencedirect.com/science/article/pii/S0962629810000132.

Rogers, Andrei, and James Raymer. 1998. "The spatial focus of US interstate migration flows". *International Journal of Population Geography* 4 (1): 63–80. ISSN: 1099-1220. doi:10.1002/(sici)1099-1220(199803)4:1<63::aid-ijpg87>3.0.co;2-u. http://dx.doi.org/10.1002/(SICI)1099-1220(199803)4:1%3C63::AID-IJPG87%3E3.0.CO;2-U.

Rogers, Andrei, James Raymer, and Frans Willekens. 2002. "Capturing the Age and Spatial Structures of Migration". *Environment and Planning A* 34, no. 2 (): 341–359. ISSN: 0308-

518X, 1472-3409, visited on 05/13/2016. doi:10.1068/a33226. http://epn.sagepub.com/lookup/doi/10.1068/a33226.

Roseman, Curtis C. 1983. "Labor force migration, non-labor force migration, and non-employment reasons for migration". *Socio-Economic Planning Sciences* 17, **numbers** 5-6 (): 303–312. ISSN: 00380121, visited on 04/09/2021. doi:10.1016/0038-0121(83)90037-X. https://linkinghub.elsevier.com/retrieve/pii/003801218390037X.

Roseman, Curtis C, and Kevin E McHugh. 1982. "Metropolitan areas as redistributors of population". Publisher: Taylor & Francis, *Urban Geography* 3 (1): 22–33. ISSN: 0272-3638.

Rupasingha, Anil, Yongzheng Liu, and Mark Partridge. 2015. "Rural Bound: Determinants of Metro to Non-Metro Migration in the United States". *American Journal of Agricultural Economics* 97, no. 3 (): 680–700. ISSN: 0002-9092, 1467-8276, visited on 06/30/2021. doi:10.1093/ajae/aau113. https://onlinelibrary.wiley.com/doi/abs/10.1093/ajae/aau113.

Sasser, Alicia C. 2010. "Voting with their feet: Relative economic conditions and state migration patterns". *Regional Science and Urban Economics* 40, **numbers** 2-3 (): 122–135. ISSN: 01660462, visited on 05/30/2018. doi:10.1016/j.regsciurbeco.2010.02.001. http://linkinghub.elsevier.com/retrieve/pii/S0166046210000049.

Scott, Allen J. 2010. "Jobs or amenities? Destination choices of migrant engineers in the USA*: Migrant engineers". *Papers in Regional Science* 89, no. 1 (): 43–63. ISSN: 10568190, 14355957, visited on 07/09/2021. doi:10.1111/j.1435-5957.2009.00263.x. https://onlinelibrary.wiley.com/doi/10.1111/j.1435-5957.2009.00263.x.

Senior, Martyn L. 1979. "From gravity modelling to entropy maximizing: a pedagogic guide". *Progress in Geography* 3, no. 2 (): 175–210. ISSN: 0556-1892, visited on 06/03/2019. doi:10.1177/030913257900300218. https://doi.org/10.1177/030913257900300218.

Shumway, J Matthew, and Samuel Otterstrom. 2010. "US regional income change and migration: 1995-2004". *Population, Space and Place* 16 (6): 483–497. ISSN: 1544-8452. doi:10.1002/psp.566. http://dx.doi.org/10.1002/psp.566.

Shumway, J. Matthew, Samuel Otterstrom, and Sonya Glavac. 2014. "Environmental Hazards as Disamenities: Selective Migration and Income Change in the United States from 2000-2010". *Annals of the Association of American Geographers* 104, no. 2 (): 280–291. ISSN: 0004-5608, 1467-8306, visited on 05/13/2016. doi:10.1080/00045608.2013.873322. http://www.tandfonline.com/doi/abs/10.1080/00045608.2013.873322.

Sidorov, Grigori, et al. 2014. "Soft similarity and soft cosine measure: Similarity of features in vector space model". Publisher: Centro de Investigación en Computación, IPN, *Computación y Sistemas* 18 (3): 491–504. ISSN: 1405-5546.

Smith, Lilly. 2020. *'Zoom towns' are exploding in the West*. Visited on 06/16/2021. https://www.fastcompany.com/90564796/zoom-towns-are-exploding-in-the-west.

SOI Tax Stats - Migration Data — Internal Revenue Service. 2018. https://www.irs.gov/statistics/soi-tax-stats-migration-data.

Stewart, John Q. 1950. "The development of social physics". Publisher: American Association of Physics Teachers, *American Journal of Physics* 18 (5): 239–253. ISSN: 0002-9505.

Stockdale, Aileen, Marsaili MacLeod, and Lorna Philip. 2013. "Connected Life Courses: Influences on and Experiences of 'Midlife' In-Migration to Rural Areas". *Population, Space and Place* 19, no. 3 (): 239–257. ISSN: 15448444, visited on 04/14/2021. doi:10.1002/psp.1709. http://doi.wiley.com/10.1002/psp.1709.

Stoll, Cliff. 2005. *The cuckoo's egg: tracking a spy through the maze of computer espionage*. Simon / Schuster. ISBN: 1-4165-0778-7.

Stouffer, Samuel A. 1960. "INTERVENING OPPORTUNITIES AND COMPETING MIGRANTS". Publisher: John Wiley & Sons, Ltd, *Journal of Regional Science* 2, no. 1 (): 1–

26. ISSN: 0022-4146, visited on 05/15/2020. doi:10.1111/j.1467-9787.1960.tb00832.x. https://doi.org/10.1111/j.1467-9787.1960.tb00832.x.

Tarver, James D., and R. Douglas McLeod. 1973. "A Test and Modification of Zipf's Hypothesis for Predicting Interstate Migration". *Demography* 10 (2): 259–275. ISSN: 00703370, 15337790. doi:10.2307/2060817. http://www.jstor.org.offcampus.lib.washington.edu/stable/2060817.

Taylor, Laura. 2011. "No boundaries: exurbia and the study of contemporary urban dispersion". *GeoJournal* 76, no. 4 (): 323–339. ISSN: 0343-2521, 1572-9893, visited on 05/30/2021. doi:10.1007/s10708-009-9300-y. http://link.springer.com/10.1007/s10708-009-9300-y.

Thapa, Mukund Narain, and George B. (George Bernard) Dantzig. 1997. *Linear programming*. Ed. by Mukund Narain Thapa. New York: New York : Springer.

Thompson, Derek. 2020. *The Coronavirus Is Creating a Huge, Stressful Experiment in Working From Home*. Section: Ideas. Visited on 06/16/2021. https://www.theatlantic.com/ideas/archive/2020/03/coronavirus-creating-huge-stressful-experiment-working-home/607945/.

Tobler, W R. 1970. "A Computer Movie Simulating Urban Growth in the Detroit Region". *Economic Geography* 46:234–240. ISSN: 00130095. http://www.jstor.org/stable/143141.

Tobler, Waldo. 1983. "An alternative formulation for spatial-interaction modeling". *Environment and Planning A* 15 (5): 693–703. ISSN: 0308-518X.

Troutt, Marvin D. 2006. "Regression, 10k Rule of Thumb for". In *Encyclopedia of Statistical Sciences*. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/0471667196.ess6045.pub2. American Cancer Society. ISBN: 978-0-471-66719-3. doi:10.1002/0471667196.ess6045.pub2. https://onlinelibrary.wiley.com/doi/abs/10.1002/0471667196.ess6045.pub2.

Tufte, Edward R, Nora Hillman Goeler, and Richard Benson. 1990. *Envisioning information*. Vol. 126. Graphics press Cheshire, CT.

US Census Bureau. 1990. *METROPOLITAN AREAS AND COMPONENTS, 1990 WITH FIPS CODES*. Visited on 04/07/2020. https://www2.census.gov/programs-surveys/metro-micro/geographies/reference-files/1990/historical-delineation-files/90mfips.txt.

— . 1993. *METROPOLITAN AREAS AND COMPONENTS, 1993, WITH FIPS CODES*. https://www2.census.gov/programs-surveys/metro-micro/geographies/reference-files/1993/historical-delineation-files/93mfips.txt.

— . 1999. *METROPOLITAN AREAS AND COMPONENTS, 1999, WITH FIPS CODES*. https://www2.census.gov/programs-surveys/metro-micro/geographies/reference-files/1999/historical-delineation-files/99mfips.txt.

— . 2003. *METROPOLITAN AND MICROPOLITAN STATISTICAL AREAS AND COMPONENTS, December 2003, WITH CODES*. https://www2.census.gov/programs-surveys/metro-micro/geographies/reference-files/2003/historical-delineation-files/0312mfips.txt.

— . 2004. *METROPOLITAN AND MICROPOLITAN STATISTICAL AREAS AND COMPONENTS, November 2004, WITH CODES*. Visited on 04/07/2020. https://www2.census.gov/programs-surveys/metro-micro/geographies/reference-files/2004/historical-delineation-files/list1.txt.

— . 2005. *METROPOLITAN AND MICROPOLITAN STATISTICAL AREAS AND COMPONENTS, December 2005, WITH CODES*. Visited on 04/07/2020. https://www2.census.gov/programs-surveys/metro-micro/geographies/reference-files/2005/historical-delineation-files/list1.txt.

— . 2006. *METROPOLITAN AND MICROPOLITAN STATISTICAL AREAS AND COMPONENTS, December 2006, WITH CODES*. Visited on 04/07/2020. https://www2.

census.gov/programs-surveys/metro-micro/geographies/reference-files/2006/historical-delineation-files/list1.txt.

— . 2007. *METROPOLITAN AND MICROPOLITAN STATISTICAL AREAS AND COMPONENTS, November 2007, WITH CODES*. Visited on 04/07/2020. https://www2.census.gov/programs-surveys/metro-micro/geographies/reference-files/2007/historical-delineation-files/list1.txt.

— . 2009. *METROPOLITAN AND MICROPOLITAN STATISTICAL AREAS AND COMPONENTS, November 2008, WITH CODES*. https://www2.census.gov/programs-surveys/metro-micro/geographies/reference-files/2008/historical-delineation-files/list1.txt.

— . 2010. *METROPOLITAN AND MICROPOLITAN STATISTICAL AREAS AND COMPONENTS, December 2009, WITH CODES*. Visited on 04/07/2020. https://www2.census.gov/programs-surveys/metro-micro/geographies/reference-files/2009/historical-delineation-files/list1.txt.

— . 2013. *List 1. CORE BASED STATISTICAL AREAS (CBSAs) AND COMBINED STATISTICAL AREAS (CSAs), FEBRUARY 2013*. Visited on 04/07/2020. https://www2.census.gov/programs-surveys/metro-micro/geographies/reference-files/2013/delineation-files/list1.xls.

— . 2016. *List 1. CORE BASED STATISTICAL AREAS (CBSAs), METROPOLITAN DIVISIONS, AND COMBINED STATISTICAL AREAS (CSAs), JULY 2015*. Visited on 04/07/2020. https://www2.census.gov/programs-surveys/metro-micro/geographies/reference-files/2015/delineation-files/list1.xls.

— . 2019. *Changes to Counties and County Equivalent Entities: 1970-Present*. https://www.census.gov/programs-surveys/geography/technical-documentation/county-changes.html.

US Energy Information Administration. 2021. *Degree-days - U.S. Energy Information Administration (EIA)*. Visited on 06/28/2021. https://www.eia.gov/energyexplained/units-and-calculators/degree-days.php.

USDA Economic Research Service. 2020. *Rural Urban Continuum Codes*. Visited on 07/04/2021. https://www.ers.usda.gov/data-products/rural-urban-continuum-codes/documentation/.

Van Rossum, Guido, and Fred L Drake Jr. 1995. *Python tutorial*. Centrum voor Wiskunde en Informatica Amsterdam.

Ver Hoef, Jay M, and Peter L Boveng. 2007. "Quasi-Poisson vs. negative binomial regression: how should we model overdispersed count data?" Publisher: Wiley Online Library, *Ecology* 88 (11): 2766–2772. ISSN: 1939-9170.

Vining, D R, and A Strauss. 1977. "A Demonstration That the Current Deconcentration of Population in the United States is a Clean Break with the Past". *Environment and Planning A: Economy and Space* 9, no. 7 (): 751–758. ISSN: 0308-518X, 1472-3409, visited on 06/23/2021. doi:10.1068/a090751. http://journals.sagepub.com/doi/10.1068/a090751.

Vose, Russell S., et al. 2014. *NOAA's Climate Divisional Database (nCLIMDIV)*. Type: dataset. Visited on 06/28/2021. doi:10.7289/V5M32STR. https://data.nodc.noaa.gov/cgi-bin/iso?id=gov.noaa.ncdc:C00005.

Wall, Tamara, and Christiane Von Reichert. 2013. "Divorce as an Influence in Return Migration to Rural Areas: Divorce and Rural Return Migration". *Population, Space and Place* 19, no. 3 (): 350–363. ISSN: 15448444, visited on 04/14/2021. doi:10.1002/psp.1719. http://doi.wiley.com/10.1002/psp.1719.

Ware, Colin. 2019. *Information visualization: perception for design*. Morgan Kaufmann. ISBN: 0-12-812876-3.

Williams, James D, and Andrew J Sofranko. 1979. "Motivations for the inmigration component of population turnaround in nonmetropolitan areas". Publisher: Springer, *Demography* 16 (2): 239–255. ISSN: 1533-7790.

Wilson, Alan Geoffrey. 1971. "A family of spatial interaction models, and associated developments". *Environment and Planning A* 3 (1): 1–32. ISSN: 0308-518X.

Withers, Suzanne Davies, and William A V Clark. 2006. "Housing costs and the geography of family migration outcomes". *Population, Space and Place* 12 (4): 273–289. ISSN: 1544-8452. doi:10.1002/psp.415. http://dx.doi.org/10.1002/psp.415.

Withers, Suzanne Davies, William A V Clark, and Tricia Ruiz. 2008. "Demographic variation in housing cost adjustments with US family migration". *Population, Space and Place* 14 (4): 305–325. ISSN: 1544-8452. doi:10.1002/psp.503. http://dx.doi.org/10.1002/psp.503.

Wright, Richard, and Mark Ellis. 2019. "Where science, technology, engineering, and mathematics (STEM) graduates move: Human capital, employment patterns, and interstate migration in the United States". *Population, Space and Place* 25, no. 4 (): e2224. ISSN: 1544-8444, 1544-8452, visited on 07/08/2021. doi:10.1002/psp.2224. https://onlinelibrary.wiley.com/doi/10.1002/psp.2224.

Yano, Keiji, et al. 2003. "A comparison of migration behaviour in Japan and Britain using spatial interaction models". *International Journal of Population Geography* 9 (5): 419–431. ISSN: 1099-1220. doi:10.1002/ijpg.297. http://dx.doi.org/10.1002/ijpg.297.

Young, Ernest Charles. 1924. *The movement of farm population*. Vol. 426. Cornell University Agricultural Experiment Station.

Zagheni, Emilio, et al. 2014. "Inferring international and internal migration patterns from twitter data", 439–444.

Zelinsky, Wilbur. 1971. "The hypothesis of the mobility transition". *Geographical Review* 61:219–249.

Zhang, Sumei, and Jean-Michel Guldmann. 2015. "A Regression-constrained Optimization Approach to Estimating Suppressed Information using Time-series Data: Application to County Business Patterns 1999 - 2006". *International Regional Science Review* 38, no. 2 (): 119–150. ISSN: 0160-0176, 1552-6925, visited on 06/04/2021. doi:10.1177/0160017613501866. http://journals.sagepub.com/doi/10.1177/0160017613501866.

Zipf, George Kingsley. 1946. "The P1 P2/D Hypothesis: On the Intercity Movement of Persons". *American Sociological Review* 11 (6): 677–686. ISSN: 00031224. doi:10.2307/2087063. http://www.jstor.org.offcampus.lib.washington.edu/stable/2087063.

— . 1949. *Human behavior and the principle of least effort*. Cambridge, Mass.: Cambridge, Mass., Addison-Wesley Press.

Zonghao Gu, Edward Rothberg, and Robert Bixby. 2019. *Gurobi*.

Appendix A

COUNTS AND RATES OF FOREIGN MIGRATION, 1990-2015

Figure A.1: Counts and rates of foreign migration, 1990-2015

Appendix B

REPORTED AND GENERATED RECORDS BY YEAR, 1990-2015

Table B.1: Reported and generated records by year, 1990-2015

		Incoming			Outgoing		
Tax Year	Flow Type	Reported	Generated	Total	Reported	Generated	Total
1990	same state	30,698	-	30,698	30,698	-	30,698
	diff. state	43,763	-	43,763	43,763	-	43,763
1991	same state	30,922	-	30,922	30,922	-	30,922
	diff. state	43,568	-	43,568	43,568	-	43,568
1992	same state	31,137	-	31,137	31,137	-	31,137
	diff. state	43,982	-	43,982	43,982	-	43,982
1993	same state	31,306	-	31,306	31,306	-	31,306
	diff. state	43,695	-	43,695	43,695	-	43,695
1994	same state	31,815	-	31,815	31,815	-	31,815
	diff. state	45,372	-	45,372	45,372	-	45,372
1995	same state	31,809	67	31,876	31,796	80	31,876
	diff. state	44,597	2	44,599	44,595	4	44,599
1996	same state	32,589	65	32,654	32,593	61	32,654
	diff. state	45,671	5	45,676	45,673	3	45,676
1997	same state	32,731	76	32,807	32,694	113	32,807
	diff. state	46,191	5	46,196	46,183	13	46,196
1998	same state	33,368	61	33,429	33,363	66	33,429
	diff. state	46,835	3	46,838	46,835	3	46,838
1999	same state	33,644	68	33,712	33,654	58	33,712
	diff. state	47,232	7	47,239	47,236	3	47,239
2000	same state	33,732	62	33,794	33,716	78	33,794
	diff. state	47,839	6	47,845	47,842	3	47,845
2001	same state	33,795	67	33,862	33,791	71	33,862
	diff. state	47,293	52	47,345	47,298	47	47,345
2002	same state	33,312	87	33,399	33,303	96	33,399
	diff. state	45,898	29	45,927	45,868	59	45,927
2003	same state	33,274	66	33,340	33,265	75	33,340
	diff. state	46,042	5	46,047	46,045	2	46,047
2004	same state	33,586	65	33,651	33,573	78	33,651
	diff. state	48,068	-	48,068	48,068	-	48,068
2005	same state	34,671	64	34,735	34,654	81	34,735
	diff. state	50,646	-	50,646	50,646	-	50,646
2006	same state	33,983	56	34,039	33,957	82	34,039
	diff. state	49,672	-	49,672	49,672	-	49,672
2007	same state	34,935	70	35,005	34,917	88	35,005
	diff. state	50,634	-	50,634	50,634	-	50,634
2008	same state	34,337	63	34,400	34,319	81	34,400
	diff. state	48,765	1	48,766	48,766	-	48,766
2009	same state	32,915	65	32,980	32,892	88	32,980
	diff. state	45,402	-	45,402	45,402	-	45,402
2010	same state	33,300	58	33,358	33,284	74	33,358
	diff. state	47,833	1	47,834	47,833	1	47,834
2011	same state	35,148	-	35,148	35,148	-	35,148
	diff. state	54,379	-	54,379	54,379	-	54,379
2012	same state	35,431	-	35,431	35,431	-	35,431
	diff. state	56,028	-	56,028	56,028	-	56,028
2013	same state	21,528	-	21,528	21,528	-	21,528
	diff. state	23,920	-	23,920	23,920	-	23,920
2014	same state	18,439	-	18,439	18,439	-	18,439
	diff. state	17,562	-	17,562	17,562	-	17,562
2015	same state	21,906	-	21,906	21,906	-	21,906
	diff. state	24,983	-	24,983	24,983	-	24,983
	Totals	1,980,181	1,176	1,981,357	1,979,949	1,408	1,981,357

Appendix C

INTRASTATE MIGRATION SOLUTIONS, NEVADA, 2002

Table C.1: Solutions for intrastate migration, Nevada, 2002, upper bounds of 6 and 7

Maximum Upper Bound of 6

County FIPS	32001	32003	32005	32007	32009	32011	32013	32015	32017	32019	32021	32023	32027	32029	32031	32033	32510	Reported	Remainder	Total
32001		23	11	15	6	6	6	3		62	15	17			105		15	263	21	284
32003	29		73	105	6	6	13	12	26	32	6	397	12	1	629	44	94	1,466	19	1,485
32005	6	46		6	2	6	6	5		63					181		273	563	31	594
32007	6	68	6			21	34	34	1	11		24			52	17	12	273	13	286
32009	6	6	6																18	18
32011	6	6	6	5											4				27	27
32013	6	17	6	27				17	4	11	6	2	13		44			129	24	153
32015	6	6	2	14			14			4		1			6			28	25	53
32017	1	50										6			6			50	13	63
32019	79	54	115	17			6	6			6	21	6	15	503	6	362	1,166	30	1,196
32021	11	6		6			6			12		6			37		1	60	25	85
32023	6	657	3	12			6	6		15	6		6	6	32	11	12	739	39	778
32027	1	6		4			6	6		3		6			15		6	15	38	53
32029		6					2	1		3		6			256	6	16	272	24	296
32031	151	588	232	126		6	106	22	6	406	24	42	27	214		25	445	2,408	12	2,420
32033		50		13						5		14		4	12		6	89	15	104
32510	20	86	401	14				3		273	6	6	6	16	329	13		1,152	21	1,173
Reported	290	1,639	832	343		21	167	85	26	885	39	515	52	245	2,195	110	1,229	8,673		
Remainder	44	36	29	21	14	24	38	30	11	15	30	33	18	11	16	12	13		395	
Total	334	1,675	861	364	14	45	205	115	37	900	69	548	70	256	2,211	122	1,242			9,068

87 reported county to county pairs
79 estimated county to county pairs

Maximum Upper Bound of 7

County FIPS	32001	32003	32005	32007	32009	32011	32013	32015	32017	32019	32021	32023	32027	32029	32031	32033	32510	Reported	Remainder	Total
32001		23	11	15			7	7		62	15	17	7		105		15	263	21	284
32003	29		73	105	2	7	13	12	26	32	7	397	12	3	629	44	94	1,466	19	1,485
32005	1	46		7			7	7		63		7	2		181		273	563	31	594
32007	7	68				21	34	34		11	6	24			52	17	12	273	13	286
32009		1								1		5			6		5		18	18
32011	1	7					2			7		7			3				27	27
32013	7	17	4	27				17		11		7	13		44	5	1	129	24	153
32015	7	7	4	14			14				7							28	25	53
32017	6	50													7			50	13	63
32019	79	54	115	17			7	7			7	21	2	15	503	7	362	1,166	30	1,196
32021	11	7	7	7		3	1			12					37			60	25	85
32023	7	657	7	12		7	7	7		15	3			1	32	11	12	739	39	778
32027	7	7	7	6					4					7	15			15	38	53
32029		7		1			7	2				7			256		16	272	24	296
32031	151	588	232	126	5		106	22	7	406	24	42	27	214		25	445	2,408	12	2,420
32033	1	50		13						7		14			12		7	89	15	104
32510	20	86	401	14	7	7				273			7	16	329	13		1,152	21	1,173
Reported	290	1,639	832	343		21	167	85	26	885	39	515	52	245	2,195	110	1,229	8,673		
Remainder	44	36	29	21	14	24	38	30	11	15	30	33	18	11	16	12	13		395	
Total	334	1,675	861	364	14	45	205	115	37	900	69	548	70	256	2,211	122	1,242			9,068

87 reported county to county pairs
73 estimated county to county pairs

Table C.2: Solutions for intrastate migration, Nevada, 2002, upper bounds of 8 and 9

Maximum Upper Bound of 8

County FIPS	32001	32003	32005	32007	32009	32011	32013	32015	32017	32019	32021	32023	32027	32029	32031	32033	32510	Reported	Remainder	Total
32001		23	11	15	8	8	5			62	15	17			105		15	263	21	284
32003	29		73	105	6	8	13	12	26	32	5	397	12		629	44	94	1,466	19	1,485
32005	8	46		8		4	8	3		63					181		273	563	31	594
32007	8	68	5			21	34	34		11		24			52	17	12	273	13	286
32009	8	8	2																18	18
32011	8	8	8	3															27	27
32013	8	17	8	27				17	3	11	5		13		44			129	24	153
32015	4	8	5	14			14								8			28	25	53
32017		50	1	3								1			8			50	13	63
32019	79	54	115	17			8	8			8	21	6	15	503		362	1,166	30	1,196
32021	11	8		7			2			12		8			37			60	25	85
32023		657		12			8	8		15	8		8	7	32	11	12	739	39	778
32027		4					7	6		4		8			15	4	5	15	38	53
32029										8		8			256	8	16	272	24	296
32031	151	588	232	126		4	106	22	8	406	24	42	27	214		25	445	2,408	12	2,420
32033		50		13						3		14		4	12		8	89	15	104
32510	20	86	401	14				5		273	4	8	4	16	329	13		1,152	21	1,173
Reported	290	1,639	832	343		21	167	85	26	885	39	515	52	245	2,195	110	1,229	8,673		
Remainder	44	36	29	21	14	24	38	30	11	15	30	33	18	11	16	12	13		395	
Total	334	1,675	861	364	14	45	205	115	37	900	69	548	70	256	2,211	122	1,242			9,068

87 reported county to county pairs
64 estimated county to county pairs

Maximum Upper Bound of 9

County FIPS	32001	32003	32005	32007	32009	32011	32013	32015	32017	32019	32021	32023	32027	32029	32031	32033	32510	Reported	Remainder	Total
32001		23	11	15			9	9		62	15	17	3		105		15	263	21	284
32003	29		73	105	1	9	13	12	26	32	9	397	12		629	44	94	1,466	19	1,485
32005		46		9			9	4		63		9			181		273	563	31	594
32007	4	68				21	34	34		11	3	24	6		52	17	12	273	13	286
32009												7			7		4		18	18
32011		9		2						9		7							27	27
32013	9	17	2	27				17		11	9	1	13		44	3		129	24	153
32015	9	7		14			14				9							28	25	53
32017	4	50													9			50	13	63
32019	79	54	115	17		3	9	9				21		15	503	9	362	1,166	30	1,196
32021	11	9	9	7						12					37			60	25	85
32023	9	657	9	12	1	9	9			15				2	32	11	12	739	39	778
32027	9	9	9						2					9	15			15	38	53
32029		2		3			2	8				9			256		16	272	24	296
32031	151	588	232	126	3		106	22	9	406	24	42	27	214		25	445	2,408	12	2,420
32033		50		13						6		14			12		9	89	15	104
32510	20	86	401	14	9	3				273			9	16	329	13		1,152	21	1,173
Reported	290	1,639	832	343		21	167	85	26	885	39	515	52	245	2,195	110	1,229	8,673		
Remainder	44	36	29	21	14	24	38	30	11	15	30	33	18	11	16	12	13		395	
Total	334	1,675	861	364	14	45	205	115	37	900	69	548	70	256	2,211	122	1,242			9,068

87 reported county to county pairs
60 estimated county to county pairs

Appendix D

MODEL DIAGNOSTICS

Table D.1: R-squared training diagnostisc for models estimating within-state and interstate state returns, 1990-2015

	Within-State			Interstate		
Year	Train: Observations	Test: Observations	r-squared	Train: Observations	Test: Observations	r-squared
1990	24,559	6,139	72%	35,011	8,752	51%
1991	24,738	6,184	75%	34,855	8,713	50%
1992	24,910	6,227	76%	35,186	8,796	53%
1993	25,045	6,261	77%	34,956	8,739	51%
1994	25,452	6,363	75%	36,298	9,074	46%
1995	25,501	6,375	75%	35,676	8,918	51%
1996	26,124	6,530	72%	36,535	9,133	50%
1997	26,246	6,561	73%	36,952	9,238	49%
1998	26,744	6,685	73%	37,466	9,366	54%
1999	26,970	6,742	74%	37,784	9,445	53%
2000	27,036	6,758	74%	38,270	9,567	50%
2001	27,090	6,772	75%	37,870	9,467	49%
2002	26,720	6,679	73%	36,736	9,183	47%
2003	26,672	6,668	76%	36,834	9,208	49%
2004	26,921	6,730	74%	38,455	9,613	47%
2005	27,788	6,947	73%	40,517	10,129	38%
2006	27,232	6,807	74%	39,738	9,934	45%
2007	28,004	7,001	72%	40,508	10,126	48%
2008	27,520	6,880	75%	39,013	9,753	55%
2009	26,384	6,596	75%	36,322	9,080	55%
2010	26,687	6,671	73%	38,268	9,566	51%
2011	28,119	7,029	72%	43,504	10,875	52%
2012	28,345	7,086	72%	44,823	11,205	55%
2013	17,223	4,305	75%	19,136	4,784	52%
2014	14,752	3,687	74%	14,050	3,512	49%
2015	17,525	4,381	74%	19,987	4,996	46%

Table D.2: Within-state return model coefficients, 1990-2015

Year	Intercept	Origin Population	Destination Population	Distance	County Adjacent
1990	-3.74	0.63	0.61	-0.47	1.50
1991	-3.76	0.63	0.60	-0.47	1.54
1992	-3.81	0.64	0.60	-0.47	1.55
1993	-3.84	0.64	0.60	-0.46	1.56
1994	-3.81	0.64	0.60	-0.47	1.54
1995	-3.79	0.64	0.61	-0.48	1.52
1996	-3.83	0.64	0.61	-0.48	1.53
1997	-3.82	0.64	0.61	-0.48	1.54
1998	-3.86	0.64	0.61	-0.48	1.56
1999	-3.90	0.63	0.61	-0.47	1.59
2000	-3.86	0.63	0.61	-0.47	1.58
2001	-3.88	0.63	0.60	-0.47	1.61
2002	-3.89	0.63	0.60	-0.46	1.63
2003	-3.90	0.63	0.60	-0.46	1.65
2004	-3.92	0.63	0.60	-0.46	1.64
2005	-3.99	0.64	0.61	-0.46	1.62
2006	-3.95	0.63	0.61	-0.46	1.62
2007	-3.93	0.63	0.62	-0.47	1.62
2008	-3.91	0.62	0.62	-0.47	1.63
2009	-3.91	0.62	0.62	-0.46	1.66
2010	-3.93	0.62	0.63	-0.46	1.62
2011	-3.93	0.62	0.62	-0.47	1.67
2012	-3.94	0.62	0.62	-0.47	1.66
2013	-3.63	0.61	0.60	-0.44	1.57
2014	-3.41	0.60	0.59	-0.43	1.53
2015	-3.66	0.61	0.59	-0.44	1.58
Min:	**-3.99**	**0.60**	**0.59**	**-0.48**	**1.50**
Average:	**-3.84**	**0.63**	**0.61**	**-0.46**	**1.59**
Max:	**-3.41**	**0.64**	**0.63**	**-0.43**	**1.67**

Table D.3: Interstate return model coefficients, 1990-2015

Year	Intercept	Origin Population	Destination Population	Distance	County Adjacent	State Adjacent
1990	-7.01	0.67	0.67	-0.23	1.90	0.54
1991	-6.87	0.66	0.66	-0.23	1.93	0.53
1992	-6.80	0.67	0.65	-0.23	1.94	0.53
1993	-6.83	0.67	0.65	-0.24	1.91	0.54
1994	-6.92	0.67	0.66	-0.23	1.95	0.53
1995	-6.70	0.66	0.65	-0.24	1.94	0.52
1996	-6.80	0.65	0.66	-0.23	1.96	0.53
1997	-6.86	0.66	0.67	-0.24	1.96	0.53
1998	-6.85	0.66	0.66	-0.24	1.95	0.54
1999	-6.93	0.66	0.67	-0.24	1.93	0.53
2000	-6.98	0.66	0.67	-0.23	1.93	0.55
2001	-6.86	0.65	0.66	-0.23	1.95	0.54
2002	-6.77	0.65	0.66	-0.23	1.94	0.54
2003	-6.90	0.66	0.67	-0.24	1.94	0.55
2004	-7.08	0.67	0.67	-0.24	1.93	0.55
2005	-7.60	0.70	0.69	-0.24	1.88	0.62
2006	-7.26	0.69	0.68	-0.24	1.90	0.57
2007	-7.21	0.68	0.68	-0.24	1.92	0.56
2008	-7.03	0.67	0.67	-0.24	1.94	0.54
2009	-6.77	0.66	0.65	-0.23	1.98	0.51
2010	-7.18	0.68	0.67	-0.23	1.93	0.53
2011	-7.75	0.70	0.71	-0.24	1.94	0.53
2012	-7.81	0.70	0.71	-0.24	1.97	0.54
2013	-6.58	0.64	0.66	-0.22	1.86	0.48
2014	-6.08	0.63	0.62	-0.21	1.81	0.47
2015	-6.60	0.64	0.64	-0.21	1.87	0.49
Min:	**-7.81**	**0.63**	**0.62**	**-0.24**	**1.81**	**0.47**
Average:	**-6.96**	**0.67**	**0.67**	**-0.23**	**1.93**	**0.53**
Max:	**-6.08**	**0.70**	**0.71**	**-0.21**	**1.98**	**0.62**

Table D.4: Model diagnostics predicting the total count of returns, 1990-2015

Year	Within-State			Interstate		
	Train: Observations	Test: Observations	r-squared	Train: Observations	Test: Observations	r-squared
1990	944	236	99.7%	449	112	99.9%
1991	948	237	99.7%	451	112	99.9%
1992	972	242	99.7%	454	113	99.9%
1993	975	243	99.7%	445	111	99.9%
1994	984	245	99.6%	460	114	99.9%
1995	968	241	99.7%	449	112	99.8%
1996	978	244	99.7%	459	114	99.9%
1997	989	247	99.7%	468	116	99.9%
1998	1,019	254	99.7%	464	115	99.9%
1999	1,024	256	99.8%	479	119	99.9%
2000	1,052	262	99.7%	484	121	99.9%
2001	1,060	265	99.8%	470	117	99.9%
2002	1,076	269	99.8%	474	118	99.9%
2003	1,070	267	99.8%	481	120	99.8%
2004	1,073	268	99.7%	484	121	99.8%
2005	1,108	276	99.7%	514	128	99.9%
2006	1,072	267	99.7%	498	124	99.9%
2007	1,084	271	99.8%	502	125	99.9%
2008	1,071	267	99.7%	482	120	99.9%
2009	1,069	267	99.7%	464	116	99.8%
2010	1,053	263	99.7%	483	120	99.9%
2011	1,142	285	99.8%	530	132	99.9%
2012	1,132	283	99.8%	538	134	99.9%
2013	1,095	273	92.0%	524	131	77.4%
2014	959	239	93.2%	464	115	92.5%
2015	1,132	283	99.3%	529	132	99.7%

Table D.5: Model coefficients predicting county-to-county pairs with a value greater than zero, 1990-2015

	Variable ID													
	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1990	-6.31068	-0.07783	0.00050	0.00054	0.00003	0.00000	2.69421	2.24328	-0.09639	-0.06067	0.77846	0.78358	-0.00621	0.00061
1991	-6.46829	-0.08112	0.00054	0.00055	0.00002	0.00001	2.59810	2.35098	-0.08127	-0.06932	0.80751	0.79259	-0.00558	0.00055
1992	-6.58409	-0.08685	0.00055	0.00059	0.00008	-0.00005	2.69147	2.31902	-0.08478	-0.06200	0.75797	0.84823	-0.00595	0.00065
1993	-6.31993	-0.07823	0.00052	0.00053	0.00008	-0.00004	2.59912	2.34671	-0.08093	-0.06718	0.74331	0.80941	-0.00741	0.00074
1994	-6.42284	-0.07461	0.00050	0.00056	0.00011	-0.00006	2.66113	2.24014	-0.07801	-0.05753	0.77499	0.82938	-0.00596	0.00048
1995	-6.71279	-0.08436	0.00058	0.00061	0.00007	-0.00004	2.57235	2.40026	-0.06593	-0.06325	0.81961	0.80383	-0.00407	0.00036
1996	-6.63959	-0.08912	0.00057	0.00059	0.00005	-0.00002	2.56518	2.45368	-0.06367	-0.06795	0.83936	0.72620	-0.00464	0.00037
1997	-6.61208	-0.08670	0.00058	0.00058	0.00004	0.00000	2.53079	2.46085	-0.06157	-0.06736	0.83747	0.70059	-0.00472	0.00036
1998	-6.73786	-0.08771	0.00060	0.00059	0.00004	-0.00001	2.56301	2.54317	-0.06048	-0.07352	0.83136	0.71400	-0.00440	0.00032
1999	-6.66433	-0.08564	0.00058	0.00058	0.00003	0.00001	2.52194	2.47294	-0.05671	-0.06155	0.83434	0.72980	-0.00431	0.00030
2000	-6.64508	-0.08637	0.00060	0.00055	0.00001	0.00002	2.46544	2.56766	-0.05660	-0.06323	0.86087	0.71707	-0.00487	0.00034
2001	-6.67026	-0.07913	0.00058	0.00057	0.00004	0.00000	2.51291	2.48604	-0.06086	-0.06293	0.84798	0.72092	-0.00508	0.00044
2002	-6.85131	-0.08907	0.00061	0.00061	0.00005	-0.00001	2.65768	2.47122	-0.06317	-0.05338	0.85251	0.73205	-0.00483	0.00042
2003	-6.86018	-0.08471	0.00060	0.00060	0.00003	0.00003	2.64046	2.48849	-0.05930	-0.05099	0.86651	0.77007	-0.00478	0.00039
2004	-6.92863	-0.08040	0.00062	0.00061	0.00000	0.00004	2.61662	2.41144	-0.05412	-0.04826	0.87873	0.79382	-0.00398	0.00028
2005	-7.16214	-0.09554	0.00065	0.00067	0.00002	0.00001	2.70362	2.39294	-0.05309	-0.03842	0.86310	0.81188	-0.00356	0.00031
2006	-7.37033	-0.11008	0.00069	0.00070	0.00003	0.00000	2.72345	2.46853	-0.04613	-0.03825	0.88277	0.82484	-0.00316	0.00023
2007	-7.31697	-0.10604	0.00068	0.00068	0.00002	0.00001	2.66753	2.45588	-0.04697	-0.03957	0.88130	0.83636	-0.00299	0.00025
2008	-7.25470	-0.10303	0.00070	0.00065	-0.00001	0.00003	2.56547	2.54832	-0.03717	-0.04901	0.86470	0.78610	-0.00306	0.00017
2009	-7.31356	-0.10693	0.00071	0.00065	-0.00001	0.00004	2.58683	2.64268	-0.03887	-0.05545	0.91385	0.85622	-0.00329	0.00020
2010	-7.12626	-0.09345	0.00071	0.00061	-0.00006	0.00007	2.45020	2.58824	-0.03676	-0.05445	0.88927	0.78895	-0.00384	0.00023
2011	-7.78757	-0.13103	0.00085	0.00077	-0.00009	0.00007	2.51045	2.50928	-0.03298	-0.03950	0.88553	0.80758	-0.00246	0.00008
2012	-7.92559	-0.12977	0.00084	0.00079	-0.00008	0.00005	2.62782	2.57386	-0.03367	-0.04180	0.85886	0.75293	-0.00251	0.00012
2013	-5.79478	-0.02534	0.00037	0.00044	0.00004	-0.00001	1.39013	1.26656	-0.07821	-0.04573	1.05007	1.10133	-0.00542	0.00097
2014	-7.38458	-0.05005	0.00072	0.00075	0.00008	-0.00004	1.38653	1.34538	-0.04319	-0.04602	1.27605	1.24418	-0.00440	0.00067
2015	-9.72979	-0.08441	0.00100	0.00116	-0.00005	-0.00022	1.71109	1.78428	-0.03083	-0.04933	0.83852	0.76782	0.00440	-0.00050

ID	Description
1	Intercept
2	Upper bound
3	Number of origins
4	Number of destinations
5	Number of origins with the upper bound
6	Number of destinations with the upper bound
7	Average incoming remainder per origin
8	Average outoing remainder per destination
9	Average incoming total per origin
10	Average outgoing total per destination
11	Percent of a county's incoming flow that is in the remainder category
12	Percent of a county's outgoing flow that is in the remainder category
13	Ratio of incoming remainder to outgoing remainder
14	Ratio of incoming total to outgoing total

Table D.6: Confusion matrix predicting county-to-county pairs with a value greater than zero, 1990-2015

Year	False Negative	False Positive	True Negative	True positive
1990	0.3%	14.2%	85.2%	0.3%
1991	0.3%	13.7%	85.7%	0.3%
1992	0.3%	13.5%	85.9%	0.3%
1993	0.3%	14.1%	85.3%	0.3%
1994	0.3%	13.9%	85.5%	0.3%
1995	0.4%	13.4%	85.9%	0.3%
1996	0.3%	13.5%	85.8%	0.3%
1997	0.4%	13.6%	85.8%	0.3%
1998	0.4%	13.1%	86.2%	0.3%
1999	0.4%	13.6%	85.7%	0.3%
2000	0.4%	13.7%	85.6%	0.4%
2001	0.4%	13.4%	85.8%	0.4%
2002	0.4%	13.0%	86.2%	0.4%
2003	0.4%	13.2%	86.0%	0.4%
2004	0.4%	13.2%	86.0%	0.4%
2005	0.5%	13.0%	86.1%	0.4%
2006	0.5%	12.7%	86.4%	0.4%
2007	0.5%	12.9%	86.1%	0.5%
2008	0.4%	12.5%	86.7%	0.4%
2009	0.4%	12.7%	86.4%	0.4%
2010	0.4%	12.8%	86.4%	0.4%
2011	0.7%	12.4%	86.3%	0.6%
2012	0.6%	11.6%	87.2%	0.6%
2013	1.4%	20.5%	76.6%	1.6%
2014	0.8%	14.4%	84.0%	0.9%
2015	0.4%	7.2%	92.0%	0.4%

Appendix E

COMPARISON OF COUNTY FIPS CODES

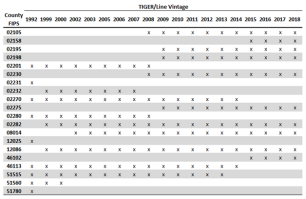
The county FIPS codes used in the IRS's county-to-county migration files should match with the county FIPS codes used by the United States Census Bureau. It should not be too surprising that there is not complete correspondence. Effort was undertaken to compare these codes for each year of data in the county-to-county migration data. As of the writing of this dissertation, I was unable to find an exhaustive list and definitive list of which counties existed in the US in which year. Possible sources include county boundary files from the National Historical Geographic Information System (NHGIS) and the TIGER files available from the US Census Bureau. Boundary files from the NHGIS were ruled out due to incomplete availability. That is, boundary files from the NHGIS were only available for 1990, 2000, and 2009 through 2017. There are two options for the 1990 and 2000 boundary files: the 2000 TIGER/Line files and the 2008 TIGER/Line files. The TIGER/Line files are datasets produced by the US Census Bureau for the purpose of the statistical administration of the US population. The files are in a proprietary format encoding vector data in a text-based file. NHGIS produced the 1990 and 2000 boundaries from the 2000 and 2008 TIGER/Line files due to increases in the accuracy of the TIGER/Line files (IPUMS 2016). While the 1990 and 2000 files feature the boundaries as relevant for the 1990 and 2000 Census Population, the 1990 and 2000 boundary files do not feature intra-decade changes as listed in the Substantial Changes To Counties and County Equivalent Entities: 1970-Present document as maintained by the US Census Bureau (US Census Bureau 2019). The non-decennial decade boundary files are therefore necessary to accurately capture and model the county-to-county flows.

There are 19 different vintages of TIGER/Line files and they correspond to the following years as depicted in Table E.1:

Table E.1: TIGER/Line vintages by decade

1990s	2000s	2010s
1992	2000	2010
1999	2002	2011
	2003	2012
	2005	2013
	2006	2014
	2007	2015
	2008	2016
	2009	2017
		2018

Table E.2: Select county membership by TIGER/Line vintage, 1992-2018



County FIPS	TIGER/Line Vintage																			
	1992	1999	2000	2002	2003	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
02105									x	x	x	x	x	x	x	x	x	x	x
02158																x	x	x	x
02195										x	x	x	x	x	x	x	x	x	x
02198										x	x	x	x	x	x	x	x	x	x
02201	x	x	x	x	x	x	x	x	x	x									
02230										x	x	x	x	x	x	x	x	x	x
02231	x																		
02232		x	x	x	x	x	x	x											
02270	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x				
02275										x	x	x	x	x	x	x	x	x	x
02280	x	x	x	x	x	x	x	x	x	x									
02282		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
08014				x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
12025	x																		
12086		x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x
46102																x	x	x	x
46113	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x				
51515	x	x	x	x	x	x	x	x	x	x	x	x	x	x	x				
51560	x	x	x																
51780	x																		

Using the Python 3.7 (Van Rossum and Drake Jr 1995) programming language and the OGR (GDAL/OGR contributors 2019) and the shapely (Gillies 2018) libraries, I extracted the county boundaries from the 1992, 1999, 2000, 2002, 2003, 2005, 2006, 2007, and 2008 TIGER/Line files and converted them to shapefiles (*ESRI Shapefile Technical Description* 1998). The TIGER/Line Shapefiles in years 2009 through 2015 feature data already in the Shapefile format. An examination of the county FIPS codes for each TIGER/Line vintage was made and the following twenty codes were found to be present in only select vintages of the TIGER/Line files. The membership is recorded in Table E.2. An empty cell indicates that a FIPS code is not present in a vintage while an "x" indicates that a FIPS code is present in each vintage. These counties were dissolved or aggregated into other larger units to make consistent counties over time.

The distances between counties comes from the US Census Bureau's TIGER/Line Files and TIGER/Line Shapefiles from the following vintages: 1992 (*1992 TIGER/Line® Files [Machine-Readable Data Files]* 1992), 1999 (*1999 TIGER/Line® Files [Machine-Readable Data Files]* 1999), 2000 (*2000 TIGER/Line® Files [Machine-Readable Data Files]* 2000), 2002 (*2002 TIGER/Line® Files [Machine-Readable Data Files]* 2002), 2003 (*2003 TIGER/Line® Files [Machine-Readable Data Files]* 2003), 2005 (*2005 TIGER/Line® Files [Machine-Readable Data Files]* 2005), 2006 (*2006 TIGER/Line® Files [Machine-Readable Data Files]* 2006), 2007 (*2007 TIGER/Line® Files [Machine-Readable Data Files]* 2007), 2008 (*2008 TIGER/Line® Shapefiles [Machine-Readable Data Files]* 2008), 2009 (*2009 TIGER/Line® Shapefiles [Machine-Readable Data Files]* 2009), 2010 (*2010 TIGER/Line®*

Shapefiles [Machine-Readable Data Files] 2010), 2011 (*2011 TIGER/Line® Shapefiles [Machine-Readable Data Files]* 2011), 2012 (*2012 TIGER/Line® Shapefiles [Machine-Readable Data Files]* 2012), 2013 (*2013 TIGER/Line® Shapefiles [Machine-Readable Data Files]* 2013), 2014 (*2014 TIGER/Line® Shapefiles [Machine-Readable Data Files]* 2014), and 2015 (*2015 TIGER/Line® Shapefiles [Machine-Readable Data Files]* 2015).

As of the writing of this dissertation, I have yet to find a specific resource featuring a manifest of counties in the US in any given year. The US Census Bureau's TIGER/Line and TIGER/Line Shapefiles provide not only the geometry for the spatial analysis but the yearly county manifest which in turn functioned as the ground truth for a yearly list of counties. I was able to compare the list of counties generated from the IRS'S county-to-county migration data to the TIGER/Line and TIGER/Line Shapefiles. The changes listed in counties and county equivalent entities since 1970 (US Census Bureau 2019) are reflected in the different vintages of the TIGER/Line data. The reason for using the TIGER/Line files and TIGER/Line Shapefiles is that the vector geometry in these files feature the full extent of the county boundaries (the full political and administrative boundaries) that enable the construction of an adjacency matrix.

After converting the different vintages of the TIGER/Line data into shapefiles, I extracted the centroid of each county for each year of data. All TIGER/Line data feature the North American Datum of 1983 (*North American Datum of 1983* 2018) as the coordinate system and therefore the generated centroids also feature the North American Datum of 1983 as the coordinate system. The centroid to centroid distance between all county-to-county pairs was calculated assuming a spherical Earth using a formula described by G. T. M. (1932) and implemented in Python. While the assumption of a spherical earth does introduce a certain amount of error into the distance calculation, it is assumed to be negligible as the distances are meant to reflect a quantifiable and specific value between any two counties. The adjacency of counties was determined using the intersect function available in the shapely software library.

These reconciled and processed lists of counties were used in conjunction with the

metropolitan data boundaries defined by the Office of Management and Budget and made available from the United States Census Bureau (US Census Bureau 1990, 1993, 1999, 2003, 2004, 2005, 2006, 2007, 2009, 2010, 2013, 2016). The metropolitan data boundaries were used to identify the metropolitan status of each county.

Appendix F

COUNTY ADJACENCY

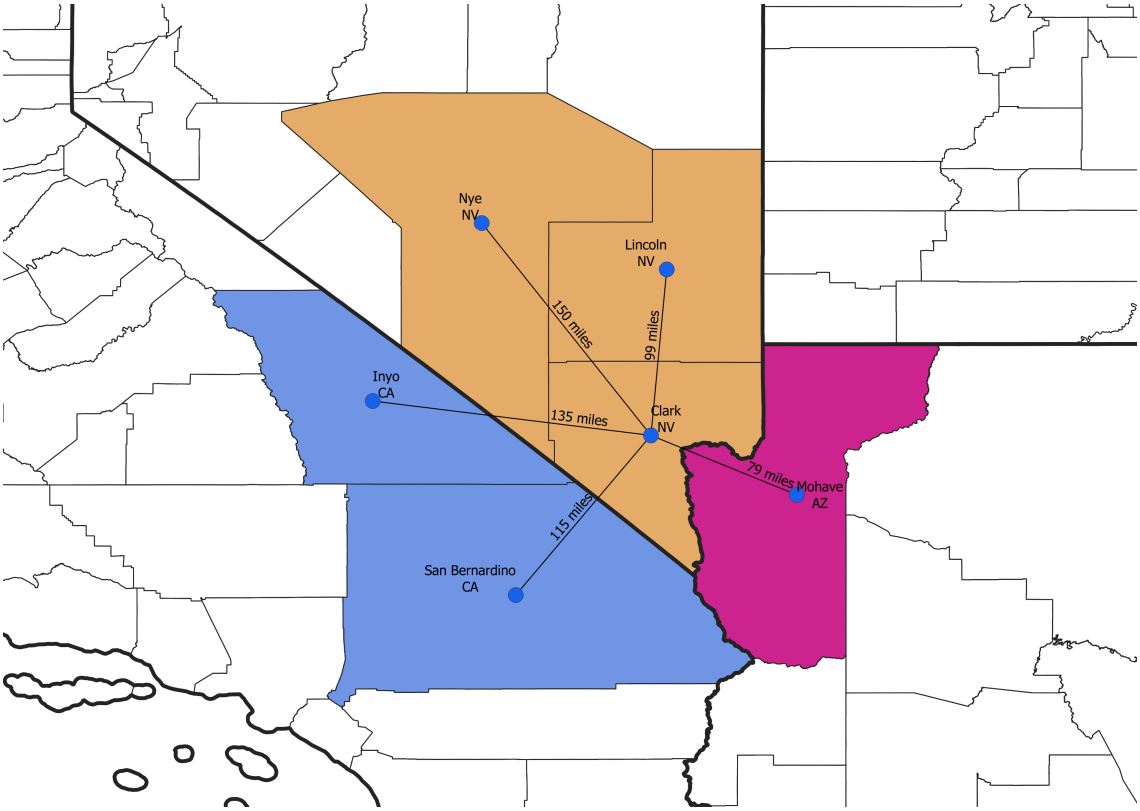


Figure F.1: Clark County, Nevada adjacency, 2018

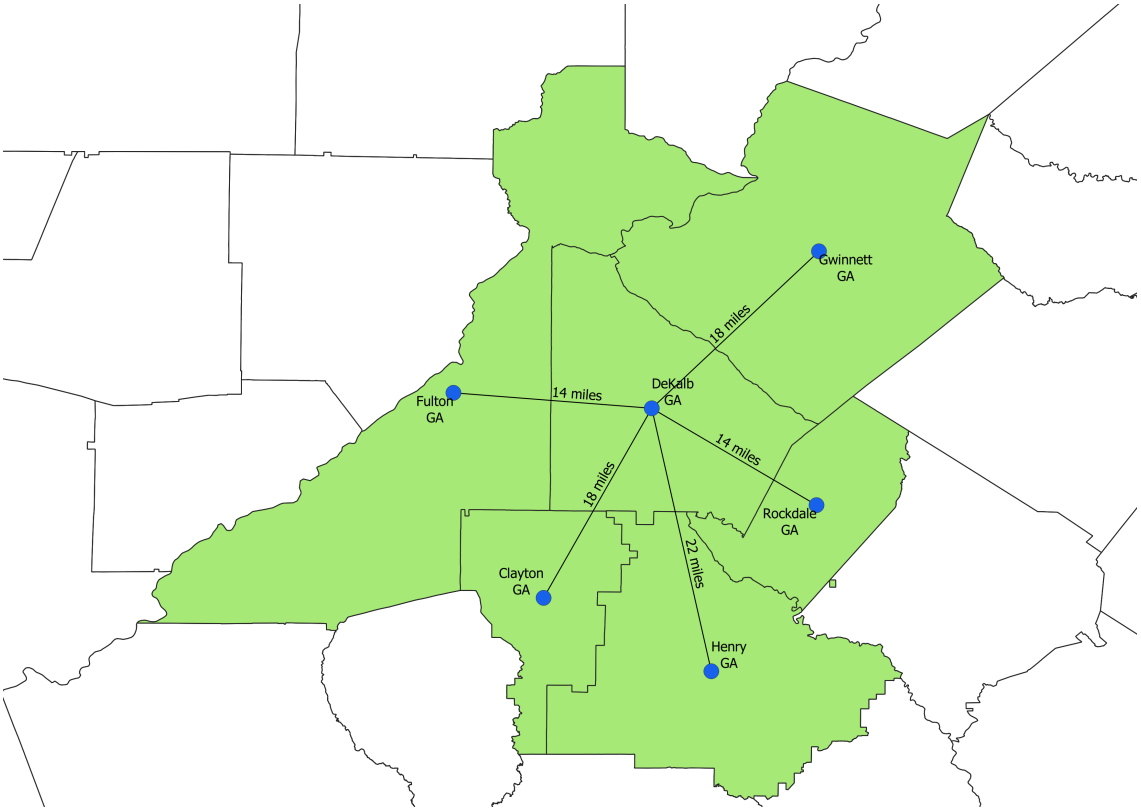


Figure F.2: DeKalb County, Georgia adjacency, 2018

VITA

Michael Babb is a computational geographer. He can be reached at babbm@uw.edu.