USING ZERO-INFLATED REGRESSION AND THE HOMOPHILY PRINCIPLE TO MODEL MIGRATION FOR POPULATION PROJECTIONS

by

Philip Morefield

A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Earth Systems and Geoinformation Sciences

Committee:

________________________________ Dr. Timothy Leslie, Dissertation Director

________________________________ Dr. Taylor Anderson, Committee Member

________________________________ Dr. Andreas Züfle, Committee Member

________________________________ Dr. Kingsley Haynes, Committee Member

________________________________ Dr. Dieter Pfoser, Department Chairperson

________________________________ Dr. Donna M. Fox, Associate Dean, Office of Student Affairs & Special Programs, College of Science

________________________________ Dr. Fernando R. Miralles-Wilhelm, Dean, College of Science

Date: __________________________ Spring Semester 2022
George Mason University
Fairfax, VA

Using Zero-Inflated Regression and the Homophily Principle to Model Migration for Population Projections

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

by

Philip Morefield
Master of Environmental Management
Portland State University, 2008
Bachelor of Science
University of Nebraska at Omaha, 2004

Director: Timothy F. Leslie, Associate Professor
Department of Geography and Geoinformation Science

Spring Semester 2022
George Mason University
Fairfax County, VA

DEDICATION

For Lucas and Daphne.

ACKNOWLEDGEMENTS

I would like to give special thanks to my advisor, Dr. Timothy Leslie, Associate Professor at George Mason University, for his patience, support, and motivation throughout my doctoral program. You instilled a sense of calm and comradery during the writing process, while also asking challenging questions that pushed me to refine my thinking numerous times. My work is undoubtedly better because of your help. I would also like to thank the members of my committee – Dr. Kingsley Haynes, Dr. Andreas Züfle, and Dr. Taylor Anderson - for sharing your valuable time with me to provide thoughtful comments and suggestions, and especially for the excellent and engaging discussions.

Finally, would also like to thank my wife, Ellie, for enthusiastically encouraging me to begin this journey, and for the many sacrifices she's made along the way.

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

LIST OF ABBREVIATIONS

M_{ij}	Observed migration from location i to location j
$\widehat{M_{ij}}$	Estimated migration from location i to location j
MAPE	Mean Absolute Percentage Error
RMSPE	Root Mean Squared Percentage Error
SSP	Shared Socioeconomic Pathway

ABSTRACT

USING ZERO-INFLATED REGRESSION AND THE HOMOPHILY PRINCIPLE TO MODEL MIGRATION FOR POPULATION PROJECTIONS

Philip Morefield, Ph.D.

George Mason University, 2022

Dissertation Director: Dr. Timothy F. Leslie

Sub-national population projections have become an essential component of policy relevant environmental assessment research. This dissertation develops 96 unique migration models by combining spatial variables, regression model types, and data sources. Overall performance is assessed with out-of-sample validation for each model using five error metrics. I find that a zero-inflated negative binomial model using modified versions of Stouffer's Intervening Opportunities and Competing Migrants calculations yielded the best overall results. Model performance was improved by conceptualizing migration according to the homophily principle and partitioning model inputs by race. After incorporating projections of births, deaths, and immigration, this new model was used to simulate migration and project county population. I find that this new model performs well compared to existing projections and I define new quantitative benchmarks for evaluating population projections going forward.

1. INTRODUCTION

Population projections inform public policy at multiple levels of governance. The U.S. Energy Information Administration prepares comprehensive energy outlooks that rely on regularly updated population projections that extend several decades into the future (U.S. Energy Information Administration, 2017). The U.S. Forest Service uses county-level population projections to asses future demands for natural resources decades in advance (Zarnoch et al., 2010). Recent research has shown that population growth over the next 20 to 30 years could significantly increase local electricity demand in the southern United States, particularly when interactions with climate change are considered (Allen et al., 2016). Other research has demonstrated that detailed population projections enhance assessments of future human vulnerability to sea-level rise (Hardy & Hauer, 2018; Hauer et al., 2015, 2016).

The U.S. Census Bureau population projections have been limited to national totals since 2005 and lack spatial detail needed to inform decision making. Furthermore, the most recent Census projections extend only to the year 2060 (U.S. Census Bureau, 2014b), limiting their utility for applications like climate change impacts assessments, which frequently seek to estimate exposure and damages into the latter part of the century. The underlying reasons for the reduced scope of the U.S. Census projections – the set produced in 2000 extend to 2100 – are not given, but the change may be attributable to technical challenges, available resources, or the potential for politicization of results.

Other population projections have been produced at the national (Azose et al., 2016), state (Jiang et al., 2020; University of Virginia, 2016), Commuting Zone (Hildner et al., 2015), county (Bierwagen et al., 2010; Hauer, 2019; U.S. Environmental Protection Agency, 2016; Wear & Prestemon, 2019) and sub-county levels (Jones & O'Neill, 2013; McKee et al., 2015), however many of these efforts lack either spatial or demographic detail needed to inform more disaggregated research and policy making efforts. The health impacts of climate change, for example, are highly place-specific, and vary greatly by age and race (U.S. Global Change Research Program, 2015), however the only existing population projection effort that includes detailed demographic information at the county level is Hauer (2019). Moreover, both the number and proportion of Americans age 65 and over is expected to increase considerably by 2050 (Ortman et al., 2014). This shift poses significant consequences for Federally funded social programs and health care provisioning, especially as it relates to the mobility of the elderly.

Projecting *where* these populations of interest are located is likely to be a critical component of any forward-looking assessment. For example: Hauer (2017; 2016) described the potential exposure of coastal populations to future sea level rise; an issue that may likely be exacerbated by an increasingly aged population seeking an amenity-rich life near water. Creating accurate, detailed, small-area population projections remains an extremely difficult challenge, though the development and application of population scenarios can be useful for planning purposes (Smith et al., 2013, p. 7). The State of California, for instance, developed alternative projections of future population for all 58 constituent counties in order “to provide a subjective assessment of the uncertainty of the states’ future population” (Sanstad et al., 2009).

The decisions and research questions supported by a population projection will reflect the type and level of detail of the projection methodology. For example, the projections by Jones and O’Neill (2013) simply downscale existing national-level projections of total population; neither their inputs nor their methodology provides demographic detail. This is in contrast to the state-level University of Virginia (2016) projections that provide a great deal of demographic detail, but sacrifice spatial resolution. Consistent among current sub-national population projections is the relative lack of attention paid to the process of domestic *migration*. This is striking for two reasons. First, there exists a rich literature of migration theory, modeling and quantification techniques for the U.S. which dates back more than 50 years. And second, because migration is one of the four components needed to produce a sub-national population projection (in addition to births, deaths, and immigration).

The mechanisms behind migration decisions are important when incorporating migration theory into population projections. Wilbur Zelinsky (1971) articulated a detailed theory of mobility transitions analogous in many respects to the theory of demographic transitions popular among demographers. Application of Zelinksy’s ideas could have important consequences for the spatial distribution of population in the U.S. and yet, without a dynamic, spatially explicit modeling approach, the existing sub-national population projections lack any cohesive theoretical framework of mobility. For decades most internal migration studies attribute the movements of individuals or households to economic drivers, particularly the search for higher wages, job opportunities, or housing costs. This framing of migration predominantly as a response to economic incentives is problematic in the context of long-range projections because detailed projections of regional – much less county level - economies do not exist at a temporal extent useful for this or similar research. Furthermore, it’s not even clear that migration *should* be modeled as a predominantly economic decision. Spring, Tolnay, & Crowder (2016) argue that the prominence of economic factors in migration research is largely due to the ease with which they can be quantitatively measured, and a general agreement on how those factors will affect migration at the individual and household levels. Moreover, Spring, Tolnay, & Crowder point out that a conceptual framework of migration based solely on economic factors is inconsistent with classic migration scholarship (e.g., E. Lee, 1966; Ravenstein, 1889) which emphasize the important roles of non-economic influences on migration decisions. And perhaps most importantly, a novel approach to modeling migration that does not rely on common economic variables such as wages and housing costs would be useful in the case of projections given that demographic processes tend to be more stable and predictable over time relative to economic conditions. Regardless of how migration decisions are conceptualized, there is a paucity of studies that evaluate alternative migration data sources.

2. BACKGROUND AND LITERATURE REVIEW

Migration studies and population projections represent vast literatures with surprisingly little overlap.

2.1. Population Projections

Population projections can – and frequently are – constructed by directly extrapolating the total population measured at two or more points in time. Highly sophisticated curve-fitting approaches have been used to advance this general approach, but an ultimate reliance on broad assumptions such as consistent or predictable growth rates limits the utility of projections developed in this way (Newbold, 2014, p. 37). This approach also ignores the processes that interact to yield total population, providing little opportunity to reasonably project population for sub-regions, or over time horizons that extend further into the future.

Several modern national and subnational population projections for the United States implement some form of the *cohort-component method,* which is the application of a projection methodology to subpopulations grouped by age, sex, and race (Smith et al., 2013, p. 46). The cohort-component method provides an opportunity to capture crucial differences between segments of the population. With these demographic groupings articulated in a projection framework, components of change – fertility, mortality, and migration – can be applied to portray human population dynamics more accurately. Sanderson (1998) demonstrated that incorporating spatially explicit demographic detail into population models can produce more accurate forecasts.

Numerous sub-national population projections for the United States have been produced over the last few decades. The U.S. Census Bureau published state-level population projections in 1997 and 2005, while more recent state-level projections were produced by the University of Virginia (2016) and Jiang *et al* (2020). The Urban Institute also produced some 27 population state- and commuting zone-level population projections (Martin et al., 2017). Each of these efforts used substantially different methodologies, as well as varying approaches to addressing the question of uncertainty. Recent county level projections (Hauer, 2019; U.S. Environmental Protection Agency, 2009, 2016; Wear & Prestemon, 2019) vary significantly in their methodologies, but all use scenarios of global change to guide assumptions about demographic change. Still other population projections are resolved at grids of varying resolutions (Gao, 2017; Jones & O'Neill, 2013; McKee et al., 2015; Murakami & Yamagata, 2019). There is a conspicuous absence of literature assessing and comparing the accuracy of these projections, not only across methodologies but also across spatial scales, e.g., state- versus county level accuracy.

2.2. Fertility and Mortality

Fertility, or *birth rate,* can be defined as the number of births per person (crude birth rate), per woman aged 15-44 (general fertility rate), or by parsing the total fertility rate by age or age groups (age-specific fertility rate). In the United States, the general fertility rate fluctuated greatly during the first half of the 20th century, influenced by World Wars I & II and the Great Depression. Societal changes also had a strong influence on fertility, at times. The Sexual Revolution that began in the 1960s broadened the roles and opportunities available to women, culminating with the legalization of birth control in 1972. The effect of these changes is evidenced by the marked decline in fertility post Baby Boom, which has remained stable and consistently below replacement level. Low fertility in advanced societies is driven almost entirely by intentional delay in parenthood, and generally reflects a positive economic outcome at the household level by allowing parents an opportunity to achieve financial stability before having a family (Beaujouan & Sobotka, 2017). At a more aggregate level, however, lower fertility may have a negative economic effect on a given country by limiting available human capital, narrowing the tax base, and reducing resources available to support basic public services, such as health care (R. Lee & Mason, 2014). There is some evidence that fertility rates may eventually rise in countries where economic and social development continues (Myrskylä et al., 2009).

As with fertility, mortality, or *death rate,* can be defined in terms of a crude death rate or age-specific mortality rates. The mortality rate in the United States has declined steadily since 1909. The overall trend is the result of numerous economic and technological drivers that have emerged over time. Improved medical care, water and sewerage systems, and improved nutrition make up a few of the myriad contributing factors (Weeks, 2012, Chapter 5). It is this decline in mortality – not any increase in fertility – that has driven population growth globally over the last two centuries. Weeks (2012) explains population growth simply: "It isn't that people now breed like rabbits; it's that we no longer die like flies[.]"

2.3. Migration

Migration is generally defined as the permanent or semi-permanent relocation of a person's residence across an administrative boundary. This movement maybe subject to any number of motives or obstacles and represents the most volatile component of population change. Two approaches for modeling migration appear relatively early in the literature.

The best-known and most widely used approach for simulating migration as a spatial process is the gravity model, which takes its name from Newton's Law of Gravitational Attraction. Published applications of gravity model precursors go back at least 75 years (Stewart, 1941). Work by Zipf (1946) examined the movement of passenger traffic on roads, rails and airways and produced a simple formula for estimating the intercity movement of people. Zipf's contribution was an important advancement because his formula was wholly consistent with the portions of Ravenstein's theory (1889) that related migration to population and distance. Zipf's approach would evolve to become what is now commonly accepted as the general form of the gravity model equation. In the case of predicting migration, the equation is typically given as:

$$M_{ij} = k\frac{P_i^{\beta_1}P_j^{\beta_2}}{D^{\alpha}}$$

Equation 1. The basic gravity equation of spatial interaction.

where M_{ij} is the number of migrants relocating from location i to location j; P_i and P_j are the size of the populations at the origin and destination locations, respectively; and β_1, β_2, α are k are variables to be estimated. Equation 1 represents the seminal step in the evolution of the gravity model because all terms could be estimated via linear regression. The gravity model remains one of the more popular analytic tools for estimating migration or other spatial interactions, although examples of population projections incorporating gravity model approaches are scarce.

Work by Stouffer (1960) introduced the *Intervening Opportunities and Competing Migrants*[1] model which proposed that the number of people moving between locations was attenuated not simply by distance, but rather some combination of: (i) the number opportunities that could be encountered at a shorter distance from the origin location than the destination location and (ii) the number of potential migrants located closer to the destination than the origin. Stouffer's model was shown to be supported by empirical data (Haynes et al., 1973; Miller, 1972), and found to outperform the gravity model in direct comparisons (Freymeyer & Ritchey, 1985; Galle & Taeuber, 1966; Wadycki, 1975). Variants of Stouffer's *Intervening Opportunities* (IO) and *Competing Migrants* (CM) concepts subsequently appear in numerous studies of migration and spatial interaction (Fik et al., 1992; Fik & Mulligan, 1990, 1998; Guldmann, 1999;

[1] Stouffer initially introduced the concept of Intervening Opportunities in 1940, however that approach required a notoriously complex parameter estimation (Akwawua & Pooler, 2000; Rogerson, 1986) and impractical for predictive applications (Gibson, 1975). His "redefinition" of the Intervening Opportunities calculation in this later work is far more tractable and appears more frequently in applications relevant to my dissertation.

Raphael, 1998). While there are minor variations how the *Intervening Opportunities* term is calculated, the general approach is easily explained in Figure 1:

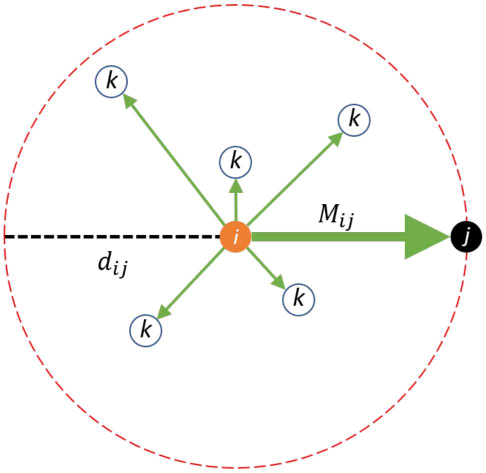

Figure 1. Schematic of the Intervening Opportunities calculation.

where *i* and *j* are the origin and destination, respectively; M_{ij} is the observed migration between *i* and *j*; D_{ij} is the distance between *i* and *j*; and intervening opportunities are denoted using *k*. The IO calculation is simply the sum of the opportunities at the intervening locations. Note that the orientation of the circle here is centered on the origin *i,* which follows Stouffer's original conceptualization of Intervening Opportunities (Stouffer, 1940). As pointed out by Wadycki (1975) this approach is more logical than Stouffer's later definition which instead uses a circle of diameter D_{ij} centered *between* the origin and destination. In short, Stouffer was suggesting in this later work that opportunities should only be considered intervening when they physically lie between a potential migrant and a potential destination.

Two other important concepts are illustrated in Figure 1. First, it is easily demonstrated that in virtually all instances the number of intervening opportunities will increase as D_{ij} increases. This is consistent with Ravenstein's original premise concerning migration flows and distance but implies that distance is merely a proxy for a cumulative number of opportunities. Second, we should expect movement between the origin and destination (M_{ij}) to attenuate as the number of intervening opportunities increases. This holds to a fundamental axiom of migration and mobility research: that an individual will preferentially select opportunities – be they employment, housing, or other – that require traversing the shortest possible distance, all other considerations being equal.

Stouffer suggests that opportunities should be defined with respect to the group of interest, and this idea was later reflected in work by Wadycki (1975) who partitioned migrations streams by occupational categories. There are two areas of literature that support partitioning migration streams by demographic characteristics. The first is earlier migration studies that describes the importance of the location of friends and family when migrants choose a destination (Gholdin, 1973; Lansing & Mueller, 1967; Price, 1971). The second comes from the field of sociology where it is generally accepted that the homophily principle – or the idea that people's social networks are largely homogenous across sociodemographic characteristics – strongly influences the makeup of social networks. As explained in a widely cited review by McPherson et al. (2001), it is race and ethnicity that play the largest role in differentiating personal social networks, followed by next by age

Like Intervening Opportunities (IO), Stouffer's Competing Migrants (CM) posits that spatial configuration of population attenuates migration flows between pairs of locations. More specifically that the possibility exists for any potential migrant – seeking employment or housing, for example – to be out-competed by individuals that live closer to the destination under consideration. The intuitive notion that the flow of information about a destination decreases with distance appears in other migration studies (Miller, 1972; Pellegrini & Fotheringham, 1999), and lends support to the idea that proximity to a destination provides a competitive advantage among prospective migrants. Although IO terms are included in quantitative migration studies more frequently than CM, support for the CM concept has been found in studies of both inter-urban migration (Galle & Taeuber, 1966) and intra-urban commuting flows (Raphael, 1998). Following Stouffer (1960) and Raphael (1998), CM is simply the sum of competing migrants that lie within a circle centered on the destination j with radius D_{ij}, as show in Figure 2:

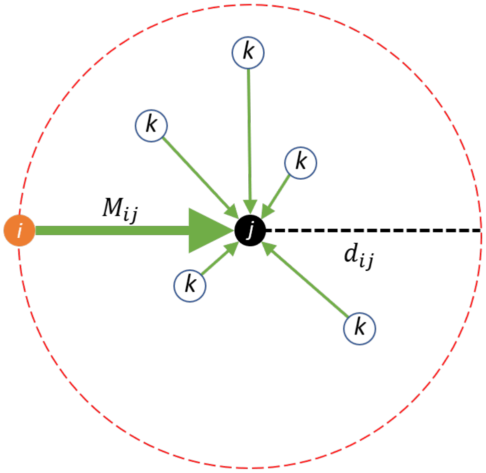

Figure 2. Schematic of the Competing Migrants calculation.

Locations more proximate to j than i are identified as k. The value of CM must be calculated for each unique origin-destination pair, at which point the term can be included in a regression analysis as predictive variable.

Another approach to incorporating spatial context into spatial interaction models, Fotheringham's *Competing Destinations* (CD) model explains destination choice as a function of human cognition and the resulting decisions to move (or not) based on a given location's position relative to other potential destinations (Fotheringham, 1983). Fotheringham argues that migrants will systematically underestimate the potential of centrally located destinations and, instead, will more likely evaluate locations that are not clustered together on the landscape. He demonstrates that a simple measure of accessibility can capture the relative position of a destination to all other destinations. As an explanatory variable in regression analyses, support for the CD model has been confirmed by others (e.g., Fik et al., 1992; Guldmann, 1999; Hu & Pooler, 2002; Lo, 1992). Each of these contributions are potentially useful for incorporating migration into population projections because they provide a sound theoretical foundation for the migration process without the need for more volatile variables such as wages, income, housing price, transportation costs, etc. that dominate the contemporary migration literature.

The measures of migration deterrence discussed in this section are well supported in the literature, however additional research that explores variants of those calculations seems worthwhile. I found no implementation of the Intervening Opportunities or Competing Migrants calculations that considered the role of distance in the influence of individual opportunities and migrants, respectively. It could be posited, for example, that further opportunities are less enticing to a potential migrant than nearby opportunities. It might similarly be the case that Competing Migrants located close to the destination are more influential than those located a further distance away.

Spatial interaction models of migration were traditionally specified using ordinary least squares (OLS) and by transforming variables to meet the assumption of normally distributed residuals. Numerous studies, however, have highlighted the deficiencies of constructing migration models using those techniques (Flowerdew & Aitkin, 1982; O’Hara & Kotze, 2010; Silva & Tenreyro, 2006). Nevertheless, estimates derived using OLS approaches are still used as benchmarks for the purposes of evaluating alternative methods (Silva & Tenreyro, 2006; Simini et al., 2012).

Over time, geographers and regional scientists began to forgo OLS in favor of maximum likelihood estimation (MLE), in particular specifying migration as a Poisson-like process (Flowerdew & Aitkin, 1982; Flowerdew & Lovett, 1988; Fotheringham & Williams, 1983). More recently, the Poisson form of spatial interaction models has gained new popularity in the analysis of international trade (Silva & Tenreyro, 2006), with some arguing that Poisson regression should be considered the "work horse" of spatial interaction modeling (Silva & Tenreyro, 2006, 2010, 2011).

In response to the work of Silva & Tenreyro, there was some debate with respect to the robustness of Poisson models when the dependent variable contains a large proportion of zeros, to which the authors responded (Silva & Tenreyro, 2011). Burger *et al* (2009) argue for the use of zero-inflated models (Lambert, 1992) in the presence of data with excess zeros, however theirs was a single analysis of bilateral trade patterns. Bohara & Krieg (1996) used zero-inflated models to analyze individual migration frequency, which is a related by substantively different issue than estimating aggregate migration. The potential for zero-inflated models to improve migration studies appears promising but remains unproven in the context of more traditional spatial interaction work.

Other recent scholarship has proposed an approach to modeling mobility and migration that conceptualizes human mobility as a process of emission and absorption, termed the *radiation model* (Simini et al., 2012). The radiation model appears to possess procedural advantages over the traditional gravity-based approach of migration modeling, as it has no parameters to estimate, making it operationally simpler than methods that require calibration. The model has been utilized to capture mobility at different temporal (i.e., diurnal and annual) and spatial (i.e., intra-urban and national) scales. This suggests a potential for development of a nested multi-scale human mobility simulation framework that uses a consistent theoretical and computational approach throughout. Ideally, a range of systematic, quantitative comparisons of the gravity and radiation models would be available to evaluate performance and applicability. While such comparisons have ostensibly been done (Masucci et al., 2013), the phenomenon under study in those examples was not migration, but rather sub-daily movement and commuting patterns. In fact, nearly all applications of the radiation model have evaluated performance only at the scale of a city and for short-term mobility as opposed to migration (e.g., Kang et al., 2015; Tizzoni et al., 2014; Wesolowski et al., 2013; Yan et al., 2014). Having said that, there are significant theoretical and methodological overlaps between migration and mobility research (e.g., Islam et al., 2021; Kavak et al., 2019; Piovani et al., 2018) which may lead to a fruitful convergence of ideas in the future.

Measuring migration is a significant challenge, and there is no single, comprehensive, authoritative data source (Smith et al., 2013, p. 117). There are three datasets that appear most often in the peer-reviewed literature. Each of these data sources provide county-to-county migration for the full extent of the U.S. but differ significantly in other aspects.

Since 1983 the Internal Revenue Service (IRS; 2016) has provided annual migration data by comparing year-to-year changes in household addresses using tax returns. No demographic detail is provided, however since 2012 the aggregate adjusted gross income for each county-to-county flow is included. In order to avoid the disclosure of personally identifiable information, all county-to-county flows of less than 20 individuals are censored.

The U.S. Census Bureau included migration questions on the “long form” questionnaire in 1990 and 2000. The 1990 Census migration data (U.S. Census Bureau, 1990) include full cross-tabulations of race, ethnicity, gender and age, making this data source entirely unique in that regard. The 2000 Census data (U.S. Census Bureau, 2000) include demographic information, although without cross-tabulation. Migration data are no longer collected as part of the decennial Census but are instead collected on a continuous basis as part of the American Community Survey (ACS; U.S. Census Bureau, 2014a).

Since 2005 the ACS data are collected annually and released on a rolling basis as an overlapping set of five-year averages, e.g., 2005-2009, 2006-2010, 2007-2011, etc. Demographic information is provided for some five-year data sets, however cross-tabulated results are not provided. The ACS also provides a 90% confidence interval for each of the county-to-county migration flows, which is a characteristic unique to this source of data.

2.4. Research Opportunities

Important gaps exist in the current literature. Many population projections do not integrate domestic migration in any capacity, limiting their utility in the face of large-scale environmental change, such as sea-level rise, which will likely affect the geography of population in the United States. Of the population projections that do consider migration explicitly, the results either lack demographic detail (U.S. Environmental Protection Agency, 2009, 2016), rely on fixed historical migration rates (Martin et al., 2017; McKee et al., 2015), or project population only at the state level (Jiang et al., 2020). A population projection approach that resolves these limitations would be a valuable contribution to the existing scholarship.

This focus on the role of migration reveals additional research opportunities. Older and perhaps outdated approaches to modeling migration such as the use of ordinary least squares and the classic gravity model persist in recent literature, despite evidence that other approaches are sounder theoretically and conceptually. Furthermore, a systematic comparison of the how migration model performance is affected by the choice of input data is conspicuously absent from the literature.

Problem Definition

This dissertation first evaluates a variety of migration models and then assesses the accuracy of a population projection that features the final migration model choice. Therefore, this dissertation requires two formal problem definitions.

First, I estimated the migration models using each of three data sources containing two to eight years of data. Each data source provides unique identifiers for the origin county *i*, destination county *j*, the number of annual migrants *M*, and the year of migration *t*. Since my goal is to predict future migration, I estimate county-to-county migration models using county population *P* for year *t* and the number of annual migrants *M* at time *t+1* between the origin county *i* and destination county *j* for all unique county pairs. Using *t* for population and *t+1* for migration avoids confounding or misspecification errors since the population of a county at year *t* is partially a function of migration for year *t.* I then validate each model by comparing the estimated number of migrants for each unique county-to-county pair $\widehat{M}_{ij}$ for year *t+1* to the observed migration $M_{ij,t+1}$.

The second major research objective is to use the newly developed and validated migration model to generate population projections. Projected gross (i.e., county to county) migration flows $\widehat{M}$ are aggregated to projected net migration M_{net} for each county for the years 2016 through 2020. So, for example, M_{net} for the year 2020 is calculated applying model coefficients to 2019 county population values, which is consistent with the estimation and validation approach described previously. The number of births, deaths and net international immigration is also projected annually by county for the years 2016 through 2020, and these values are summed (or differenced) along with M_{net} to produce an annual estimate of total population.

3. DATA

In order to contextualize and understand the research opportunities discussed in the previous section, an understanding of the data available to address the issues is needed. The idiosyncrasies of the migration data inform the research methods discussed in Chapter 4.

3.1. Migration

Despite the substantial differences in survey design and data collection methodologies, comparisons of model performance with respect to the choice of data source are completely absent from the literature. Even basic descriptive statistics of county-to-county migration datasets from the IRS, ACS and Census data sources highlights potentially important quantitative differences.

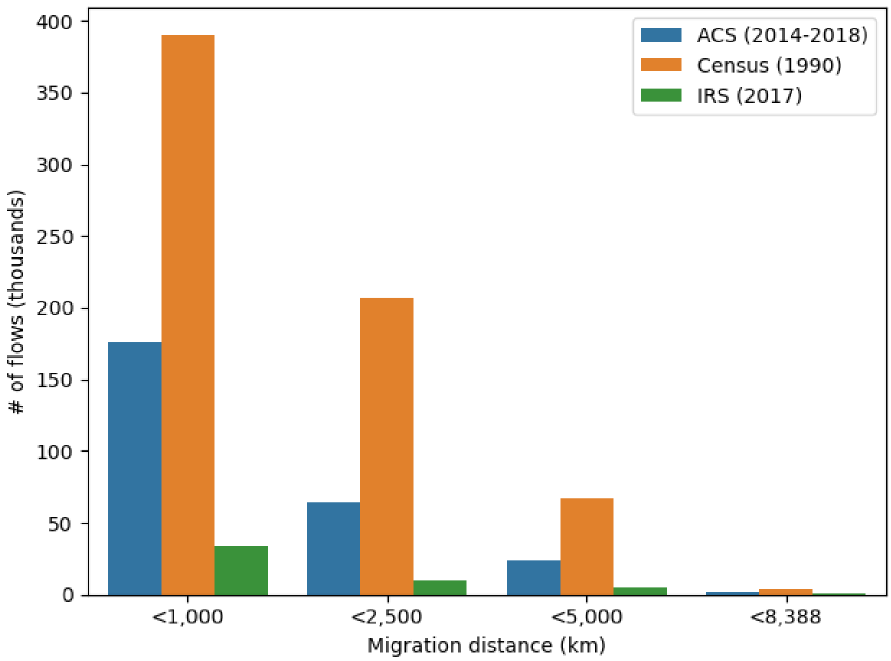

Figure 3. Distribution of migration distances for three data sources when $M_{ij} > 0$.

Figure 3 shows the distribution of migration flows by distance when migration (*M*) is less than zero. Nearly 400,000 Census data migration records covered 1,000 kilometers (km) or less, while the ACS data show less than 150,000 migration records in that distance bin. The IRS data contain nearly 40,000 migration flows that covered less than 1,000 km: a difference of roughly an order of magnitude relative to the Census. Large relative and absolute disparities are apparent in the other distance bins as well, although those differences are less apparent for moves covering more than 5,000 km Visualizing the distribution of migration flow size reveals more significant differences between the three data sources.

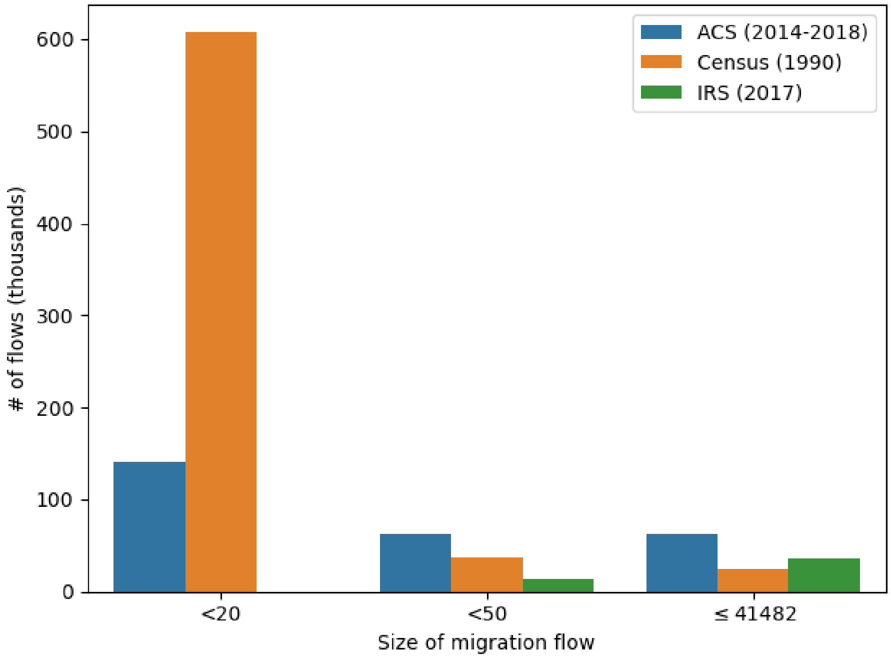

Figure 4. Distribution of migration flow size for three data sources when $M_{ij} > 0$.

Figure 4 shows that while the IRS censors migration flows of fewer than 20 individuals in the interest of privacy, the Census data from 1990 describe more than 600,000 moves in the same size bin. Both the Census and ACS data show fewer flows of in the 20-to-50 km range as migration distances increase, which is consistent with theoretical expectations. However, the IRS data the cumulative histogram of the ACS data fall between the Census and IRS, demonstrating substantial differences between the three data sources.

IRS County-to-County Migration Files

The Internal Revenue Service (IRS) disseminates annual migration counts by matching the county of residence for income tax filers in consecutive years. The IRS data are not individual migration counts *per se,* but rather county-to-county movement of tax exemptions that can be used as a proxy. Key limitations of these data have been noted in previous studies, for example the exclusion of residents that do not file tax returns. Nonetheless the annual frequency of these data makes them a unique and valuable resource for migration studies. Data are available for the years 1983 to 2017, however no demographic information on the migrants is included. The IRS suppresses all flows of less than 10 exemptions for privacy reasons through 2012 and in subsequent flows suppresses all flows of less than 20 exemptions, as shown in Figure 5.

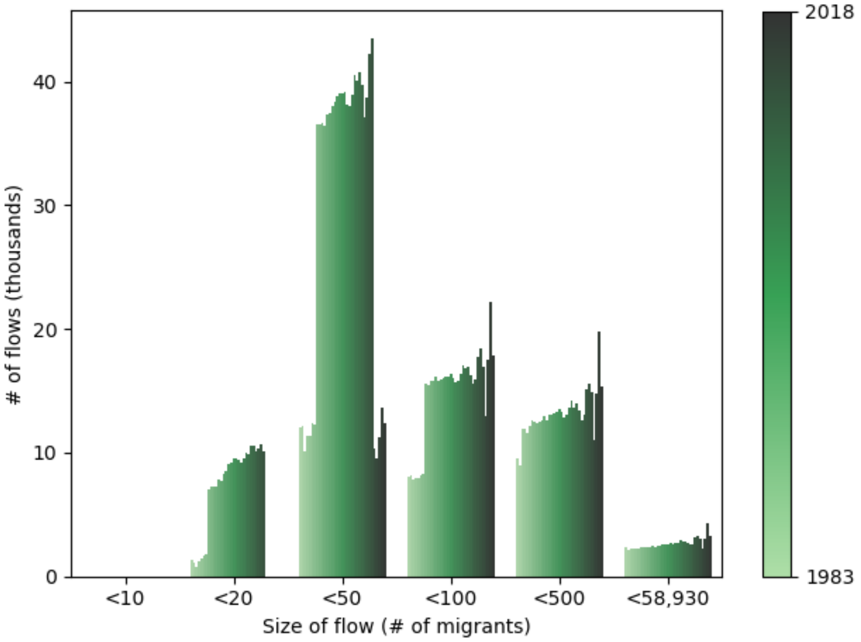

Figure 5. IRS migration flows sorted by year and binned by number of migrants.

There are also clear step changes within the <20, <50, and <100 size classes shown in Figure 5. These are likely attributable to methodological changes to the IRS tabulation procedures which are not obviously documented.

American Community Survey

The American Community Survey (ACS) measures approximately 1% of the U.S. population annually. County to county migration data are available from 2005 to 2015 and subsequent annual surveys are aggregated over rolling five-year windows. The earliest available county-to-county migration data set covers the years 2005 to 2009 and represents the average annual migration flow between counties during that period. The most recent migration data set included in this study covers the years 2011 to 2015, meaning there are a total of seven ACS migration files available covering overlapping time periods. Figure 6 shows the distribution of migration flow size for all years of the ACS data, and a discontinuity after the 2006-2010 data when the size of the flows was <10 migrants.

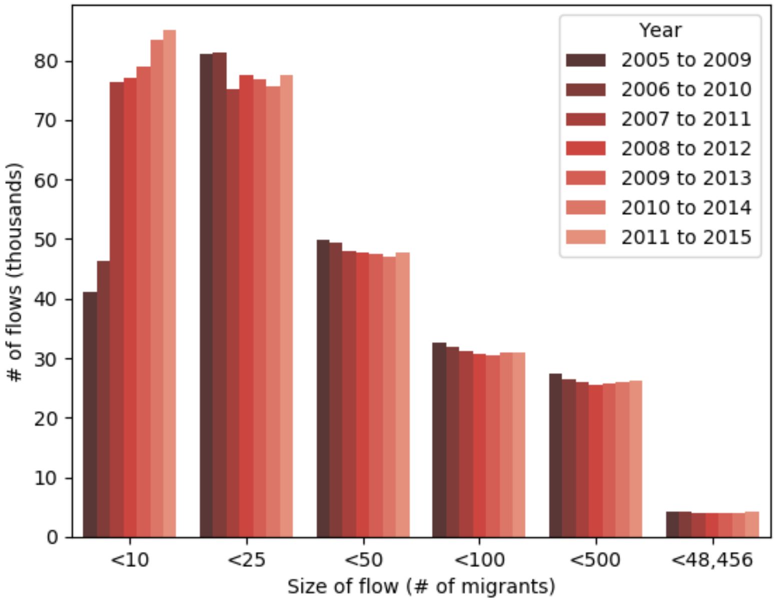

Figure 6. ACS migration flows sorted by year and binned by number of migrants.

The ACS data for 2006-2010 and 2011-2015 provide county-to-county migration flows for four race groups that were also available in the intercensal population files (White, Black, Asian, and Other), as well as Hispanic and Not-Hispanic designations. Hispanic origin is considered an ethnic trait, meaning those categories overlap the four race groups. ACS migration flows are not cross tabulated by race and Hispanic origin. Race and ethnicity are self-reported by survey respondents.

U.S. Census Bureau Enhanced Migration Files

Prior to the ACS, the sole source of demographically detailed migration data were the enhanced migration files from the U.S. Census. The data were collected as part of the decennial census in 1990 and 2000 and captured the respondent's place of residence five years prior to the actual census.

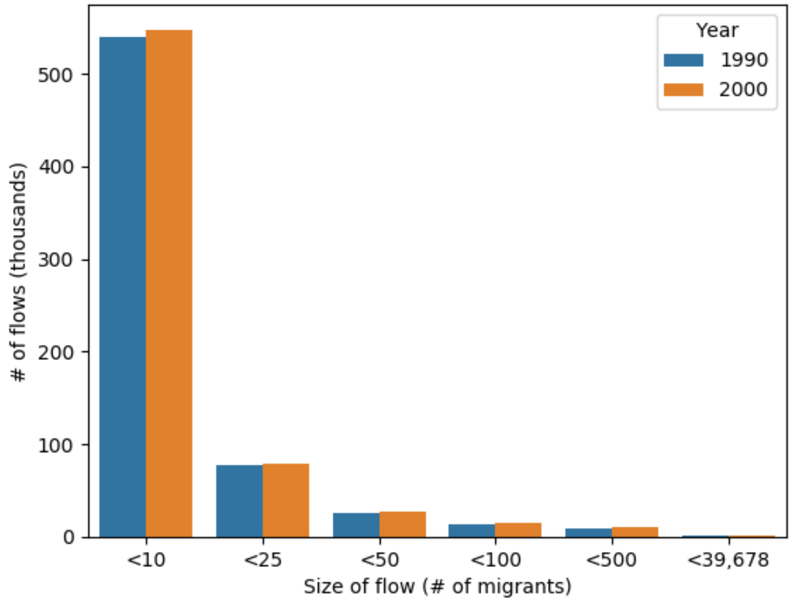

Figure 7. Census migration flows sorted by year and binned by number of migrants.

The Census migration data represent the net migration of individuals over a five-year period and not annual moves as with the IRS data, or average annual flows between counties as with the ACS. The 1990 Census enhanced migration files report county-to-county migration flows for three race groups: White, Black, and Other. The 2000 Census files replaced the Other group with two new groups: American Indian and Alaska Native and Asian and Pacific Islander. Race and ethnicity are self-reported by survey respondents.

The migration modeling approaches in sections 4.1 and 4.2 are evaluated using three sources of migration data to better understand differences in the results that may arise purely from the choice of input data. Each migration data file available from the sources described here was randomly divided into training and testing subsets comprising 80% and 20% of the complete data set, respectively. Given the apparent temporal discontinuities in Figure 5 and Figure 6, some years of IRS and ACS data were excluded. Table 1 briefly summarizes the migration data considered throughout this dissertation.

Table 1. County-to-county migration datasets used in this dissertation

	Time periods	Migration interval	Suppressions
IRS	2013 to 2017	Annual	Flows < 20
ACS	2007 to 2015	Annual (five-year average)	Small flows*
Census	1990 and 2000	Five years	Unknown†

*Flows containing only one or two people from different households, only one or two people in group quarters, or one person in group quarters and the rest from a single household are suppressed.
†Technical documentation of the long form Census questionnaire is available but makes no specific mention of migration data suppression procedures.

3.2. Population

The U.S. Census intercensal population estimates (U.S. Census Bureau, 2021) provide annual county-level population by race for all years included in this dissertation.

3.3. Labor Markets

Many studies use U.S. Census Core-based Statistical Areas (CBSA) to define labor markets. However, these geographies are limited to the most populated counties in the United States, excluding sizable portions of the country. Recent scholarship suggests that economic areas delineated using Bureau of Economic Analysis (BEA) methods provide meaningful aggregations of U.S. counties (Fowler & Jensen, 2020), and these geographic units have been featured in other migration work (Plane & Heins, 2003; Xu, 2017). Both the BEA delineations and the Economic Research Service (ERS) Commuting Zones include all U.S. counties. However, the BEA economic areas out-perform all others with respect to "containment," i.e., the population that both work and reside in the same labor market and divide only a small number of metropolitan areas (Table 2). This attribute is critical to avoiding a spatial mismatch between migration data and geographic variables. Note that use of labor markets, by any definition, does not constitute the inclusion of economic variables *per se*. These geographic definitions capture the spatial pattern of human settlement in the United States and – in addition to county boundaries – provide a defensible definition of "location" needed to both analyze and model human migration.

Table 2. Comparison of CBSA, ERS, and BEA delineations.

		Living and working in same labor market		
Delineation	Split Metropolitan Areas	Minimum	Mean	U.S. population whose commute is contained in labor market
BEA (2010)	4	85%	93%	97%
CBSA (2010)	0	36%	80%	94%
ERS (2010)	36	64%	88%	93%

Source: Fowler and Jensen (2020).

4. METHODS AND RESULTS

This dissertation research evaluates several new and existing migration models for accuracy and potential use in a full demographic model by addressing three research questions:

1. **Does the radiation model of mobility compare favorably to the traditional regression approach to spatial interactions?**
2. **Does the partitioning of migration flows by race and age improve migration model estimates?**
3. **When the best-performing migration model is combined with projections of other components of demographic change (i.e., natality, mortality, net immigration), do the resulting population projections compare favorably with existing projections?**

I address these research questions using spatial interaction modeling with a general structure similar to that of the basic gravity model (Equation 1). Instead of using distance as the denominator, however, I substitute various combinations of deterrence variables such as *Intervening Opportunities* ("origin-based") and *Competing Migrants* ("destination-based"). All deterrence variables are discussed in detail in section 4.1.1. Each model is of the same general structure:

$$M_{ij,d,t+1} = \frac{P_{i,t}^{\beta_1} \cdot P_{j,t}^{\beta_2}}{\begin{pmatrix} Origin - based \\ deterrence \\ variable \end{pmatrix}_t^{\beta_3} \cdot \begin{pmatrix} Destination - based \\ deterrence \\ variable \end{pmatrix}_t^{\beta_4}}$$

Equation 2. General form of the spatial interaction model using origin- and destination-based deterrence variables.

where $M_{ij,d,t+1}$ is the estimated number of migrants between the origin *i* and destination *j* reported in data source *d* between year *t* and *t+1*; P_i and P_j are the population at the origin and destination, respectively, in year *t* as reported by the U.S. Census Bureau; and origin- and destination-based deterrence variables (e.g., *Intervening Opportunities* and *Competing Migrants*, respectively) are one of the calculations described in section 4.1.1, also using U.S. Census Bureau data. As in Equation 1, the values of the exponents β_n are to be estimated via alternative regression approaches.

4.1. A migration model for sub-national population projections

Here I evaluate a variety of migration models using multiple accuracy metrics. Several regression model forms are estimated using combinations of variables and compared with the radiation model of mobility. The results of this section serve as a foundation for additional model development and implementation in subsequent chapters.

4.1.1. Methods

Four regression approaches were identified in the literature. The Poisson (POISSON) and ordinary least squares (OLS) models have a long record of applications in peer-reviewed migration research. Previous studies report improved migration estimates when using 119 kilometers as a breakpoint to create two separate models (Simini et al., 2012; Viboud et al., 2006). Following on those studies, I employ a segmented OLS model that separately models short- and long-distance migration using the cutoff value of 119 kilometers.

The zero-inflated Poisson (ZIP) and zero-inflated Negative Binomial (ZINB) models are relatively new but bring potential advantages to migration modeling. When considering all possible migration flows between U.S. counties, for example, it's the case that more than 90% of the observed values are zero for a given year, regardless of data source. This is a compelling justification for including a zero-inflated MLE specification, following on previous studies (Bohara & Krieg, 1996; Burger et al., 2009). The 'zero' portion of these models is specified with a complementary log-log link – as opposed to the more common logit or probit type. The complimentary log-log is typically used when a binary outcome is heavily skewed, i.e., when either absence or presence of a state or condition is extremely rare. This is indeed true of the migration data used in this study and the complimentary log-log option was found to produce better results during initial testing.

These models are used to evaluate alternative measures of deterrence – in place of Euclidean distance – using spatial measures population found throughout the literature and again repeating model estimation and validation for all three data sources. I compare the POISSON, OLS, ZIP and ZINB models to the parameter-free radiation model. All statistical models were developed and run using R.

From the literature, three measures of migration deterrence were selected for the regression analysis: Stouffer's Intervening Opportunities (S) and Competing Migrants (C) calculations, as well as Fotheringham's Competing Destinations (A) term. I also introduce two new terms which are simply distance-weighted versions of Stouffer's *S* and *C* (I assign these the letters *T* and *L*, respectively). The calculations for variables *S* and *T* are centered on the population-weighted centroid of each origin county *i* while *C, A*, and *L* are centered on the destination j. Each of these terms represents a more functional measure of deterrence to migration than simple distance and each only requires a known spatial distribution of population to be defined. Following Raphael (1998), I define the quantity of *S* and *C* as:

$$S_{ij} = \sum_{k} P_k \; \forall \; k | d_{ik} < \; d_{ij}$$

Equation 3. The Intervening Opportunities calculation.

$$C_{ij} = \sum_{k} P_k \; \forall \; k | d_{jk} < \; d_{ij}$$

Equation 4. The Competing Migrants calculation.

where P is the population at a location; i is the origin location; j is the destination location; k is all locations excluding i and j; and D_{ij} is the Euclidean distance separating the population-weighted centroids of i and j.

Following Fotheringham, Brunsdon and Charlton (2000, p. 231) I define the measure of Competing Destinations, A, as:

$$A_{ij} = \sum_k \frac{P_k}{d_{jk}}$$

Equation 5. The Competing Destinations calculation.

I also define a distance-constrained variant of Fotheringham's Competing Destinations idea as in Guldmann (1999):

$$L_{ij} = \sum_k \frac{P_k}{d_{jk}} \; \forall \; k | d_{jk} < d_{ij}$$

Equation 6. The localized Competing Destinations calculation.

Finally, I define a new migration deterrence term T_{ij} which is identical to L_{ij} except centered on the origin i:

$$T_{ij} = \sum_k \frac{P_k}{d_{ik}} \; \forall \; k | d_{ik} < d_{ij}$$

Equation 7. The calculation for a new migration deterrence variable.

The basic form of the radiation model is given as:

$$\widehat{M}_{ij} = M_i \frac{P_i P_j}{(P_i + S_{ij})(P_i + P_j + S_{ij})}$$

Equation 8. The radiation model as described in Simini et al. (2012).

where $\widehat{M}_{ij}$ is the number of migrants between the origin i and the destination j; M_i is the number of migrants originating from location i; P is the population at a location; and S_{ij} is the number of intervening opportunities between i and j as defined in Equation 3.

Out-of-sample accuracy assessments of each statistical model was conducted using 20 percent of the data for each year and data source combination that was withheld prior to model estimation. I assess the validation results using a metric described in Tofallis (2015), which I designate as Q:

$$MAPE = \frac{\sum \frac{\left|\widehat{M}_{ij} - M_{ij}\right|}{M_{ij}}}{n} \times 100$$

Equation 9. Formula for mean absolute percentage error (MAPE)

$$Q = ln\left(\frac{\widehat{M_{ij}}}{M_{ij}}\right)^2$$

Equation 10. Formula for the *Q* statistic used to assess model performance.

$$RMSPE = \sqrt{\frac{\sum \left(\frac{\widehat{M}_{ij} - M_{ij}}{M_{ij}}\right)^2 \times 100}{n}}$$

Equation 11. Formula for root mean square percentage error (RMSPE)

$$MPE = \frac{\sum \frac{\widehat{M}_{ij} - M_{ij}}{M_{ij}}}{n}$$

Equation 12. Formula for mean percentage error (MPE)

$$ME = \frac{\sum \widehat{M}_{ij} - M_{ij}}{n}$$

Equation 13. Formula for mean error (ME), calculated only when $M_{ij} = 0$.

where $\widehat{M_{\imath\jmath}}$ and M_{ij} are the predicted and actual migration, respectively, between an origin county i and destination county j. As described in Tofallis (2015), this measure of relative error is superior to more commonly used metrics, especially Mean Absolute Percentage Error (MAPE). While the MAPE statistic is easily interpretable, it is asymmetric and biased towards models that underpredict. That is, underpredictions are limited to an absolute percentage error measurement of 100% while there is obviously no theoretical limit in the case of overpredictions. Törnqvist et al. (1985) describe other mathematical limitations of MAPE and recommend an accuracy metric very similar to Q. Like all measures of relative accuracy, Q is undefined when $M = 0$, so those zero flows are validated later in this section by calculating the mean error (ME). In addition, because the IRS data do no report flows smaller than twenty individuals, the first step in the validation procedure here is to evaluate the accuracy of predictions when $M \geq 20$, followed by a second validation of ACS and Census data when $0 < M < 20$.

Finally, I report model accuracy when flows are equal to zero due the fact that more than 90% of county-to-county flows are equal to zero for all three migration data sources. To identify IRS migration flows that were likely zero, I identified origin-destination pairs with migration values of zero (or with no reported value) for all years in all data sources. This yielded more than 1.7 million unique county pairs which are assumed to have no migration interaction for the purposes of assessing accuracy. The importance of this tiered model validation approach is discussed below.

4.1.2. Results

Table 3 provides a summary of the migration deterrence variables used in combination with four types of regression model. To avoid correlated predictor variables, only combinations of one origin-centered variable and one destination-centered variable were tested. Even still, some combinations were highly correlated ($r_s > 0.6$) and were not used.

Table 3. Summary of migration deterrence variables.

Variable	*Description*	*Centered on:*	*Source(s)*	*Reference*
S_{ij}	Intervening Opportunities	Origin	Raphael, 1998; Stouffer, 1940	Figure 1 Equation 3
$S_{ij} + P_i$	Intervening Opportunities plus origin population	Origin	*	Figure 1
C_{ij}	Competing Migrants	Destination	Raphael, 1998; Stouffer, 1960	Figure 2 Equation 4
$C_{ij} + P_j$	Competing Migrants plus destination population	Destination	*	Figure 2
A_{ij}	Competing Destinations	Destination	Fotheringham et al., 2000	Equation 5
L_{ij}	Distance-weighted Competing Migrants	Destination	Guldmann, 1999	Figure 2 Equation 6
T_{ij}	Distance-weighted Intervening Opportunities	Origin	*	Figure 1 Equation 7

Note: Novel variable calculations are indicated by an asterisk (*) in the Source column.

The remaining eight origin-destination variable combinations were tested using four types of regression model estimated using three data sources. The accuracy of the resulting 96 models is shown in Table 4, Table 5 and Table 6.

Table 4. Accuracy of migration model predictions when M > 0 using ACS data.

#	Type	Var 1	Var 2	Var 3	Var 4	MAPE	RMSPE	*Q*	MPE	ME
1	OLS	P_i	P_j	S_{ij}	A_{ij}	151.1	366.4	1.32	99.2	6.23
2	OLS	P_i	P_j	$S_{ij} + P_i$	A_{ij}	149.4	355.7	1.30	97.6	6.21
3	OLS	P_i	P_j	T_{ij}	A_{ij}	151.5	364.9	1.33	99.5	6.85
4	OLS	P_i	P_j	T_{ij}	C_{ij}	150.2	359.2	1.31	98.3	6.19
5	OLS	P_i	P_j	T_{ij}	$C_{ij} + P_j$	148.2	348.5	1.29	96.5	6.18
6	OLS	P_i	P_j	S_{ij}	L_{ij}	150.2	358.6	1.31	98.3	6.16
7	OLS	P_i	P_j	$S_{ij} + P_i$	L_{ij}	148.5	350.5	1.29	96.7	6.16
8	OLS	P_i	P_j	T_{ij}	L_{ij}	150.2	357.9	1.31	98.3	6.34
9	POISSON	P_i	P_j	S_{ij}	A_{ij}	100.6	233.4	3.41	-18.5	0.51
10	POISSON	P_i	P_j	$S_{ij} + P_i$	A_{ij}	127.3	469.6	2.88	17.7	0.18
11	POISSON	P_i	P_j	T_{ij}	A_{ij}	105.0	255.3	3.70	-17.0	0.54
12	POISSON	P_i	P_j	T_{ij}	C_{ij}	99.6	224.0	3.38	-19.8	0.52
13	POISSON	P_i	P_j	T_{ij}	$C_{ij} + P_j$	120.9	376.1	2.88	9.9	0.19
14	POISSON	P_i	P_j	S_{ij}	L_{ij}	99.9	225.0	3.38	-19.5	0.52
15	POISSON	P_i	P_j	$S_{ij} + P_i$	L_{ij}	123.1	409.0	2.87	13.0	0.19
16	POISSON	P_i	P_j	T_{ij}	L_{ij}	102.0	235.4	3.56	-19.1	0.54
17	ZINB	P_i	P_j	S_{ij}	A_{ij}	131.2	774.5	2.55	23.6	0.23
18	ZINB	P_i	P_j	$S_{ij} + P_i$	A_{ij}	113.2	276.5	2.47	7.3	0.23
19	ZINB	P_i	P_j	T_{ij}	A_{ij}	132.5	777.1	2.87	20.0	0.30
20	ZINB	P_i	P_j	T_{ij}	C_{ij}	123.8	580.7	2.53	16.3	0.23
21	ZINB	P_i	P_j	T_{ij}	$C_{ij} + P_j$	111.6	285.3	2.46	5.0	0.23
22	ZINB	P_i	P_j	S_{ij}	L_{ij}	125.1	613.0	2.51	17.9	0.23
23	ZINB	P_i	P_j	$S_{ij} + P_i$	L_{ij}	112.2	286.9	2.46	5.8	0.23
24	ZINB	P_i	P_j	T_{ij}	L_{ij}	130.1	693.5	2.62	21.6	0.25

#	Type	Var 1	Var 2	Var 3	Var 4	MAPE	RMSPE	*Q*	MPE	ME
25	ZIP	P_i	P_j	S_{ij}	A_{ij}	118.1	358.1	2.50	11.6	0.23
26	ZIP	P_i	P_j	$S_{ij} + P_i$	A_{ij}	127.2	410.2	2.83	12.0	0.10
27	ZIP	P_i	P_j	T_{ij}	A_{ij}	119.6	362.9	2.84	7.6	0.30
28	ZIP	P_i	P_j	T_{ij}	C_{ij}	115.2	318.3	2.47	8.8	0.22
29	ZIP	P_i	P_j	T_{ij}	$C_{ij} + P_j$	122.1	360.2	2.78	7.3	0.10
30	ZIP	P_i	P_j	S_{ij}	L_{ij}	115.7	322.9	2.46	9.7	0.22
31	ZIP	P_i	P_j	$S_{ij} + P_i$	L_{ij}	122.7	362.0	2.74	9.0	0.10
32	ZIP	P_i	P_j	T_{ij}	L_{ij}	117.0	325.1	2.56	9.7	0.25

Var 3 and Var 4 are deterrence variables based on spatial population distribution. The best performing combination of variables for each of the four regression model types are highlighted

Table 4 displays the results of the regression models estimated using ACS migration data. Accuracy was assessed for five years of ACS data (2007-2011 through 2011-2015) and the mean value of each accuracy metrics is reported. The OLS models were the worst performers in four out of the five accuracy metrics, particularly in the measure of bias (MPE) and when predicting zero migration (ME). A Poisson model was the most accurate in two widely used metrics (MAPE and RMSPE), but there was not a clear pattern of which variable combinations should be preferred for the Poisson specification. The two zero-inflated models were competitive in across all accuracy metrics and produced the least biased (#21) and best predictor of zero migration (#26) across all ACS-based models. While there was little consensus as to the overall best variable combination, T_{ij} was a clear winner with respect to the variables centered on the origin. The best score for four of the five accuracy metrics used T_{ij} in the Var 3 column. The only exception was the ME value which was nearly identical across three different variable combinations (#26, #29, #31).

Table 5. Accuracy of migration model predictions when M > 0 using Census data.

#	Type	Var 1	Var 2	Var 3	Var 4	MAPE	RMSPE	*Q*	MPE	ME
33	OLS	P_i	P_j	S_{ij}	A_{ij}	92.5	153.7	0.87	48.4	1.23
34	OLS	P_i	P_j	$S_{ij} + P_i$	A_{ij}	91.5	152.1	0.86	47.6	1.22
35	OLS	P_i	P_j	T_{ij}	A_{ij}	94.3	156.2	0.91	50.0	1.37
36	OLS	P_i	P_j	T_{ij}	C_{ij}	92.4	153.2	0.87	48.3	1.23
37	OLS	P_i	P_j	T_{ij}	$C_{ij} + P_j$	91.1	150.0	0.85	47.3	1.22
38	OLS	P_i	P_j	S_{ij}	L_{ij}	92.4	153.3	0.87	48.3	1.22
39	OLS	P_i	P_j	$S_{ij} + P_i$	L_{ij}	91.2	150.5	0.85	47.3	1.22
40	OLS	P_i	P_j	T_{ij}	L_{ij}	92.8	153.9	0.88	48.6	1.26
41	POISSON	P_i	P_j	S_{ij}	A_{ij}	103.0	198.4	1.45	10.1	0.23
42	POISSON	P_i	P_j	$S_{ij} + P_i$	A_{ij}	102.8	212.8	1.21	2.1	0.09
43	POISSON	P_i	P_j	T_{ij}	A_{ij}	111.0	224.7	1.58	15.7	0.25
44	POISSON	P_i	P_j	T_{ij}	C_{ij}	101.6	194.5	1.44	8.4	0.23
45	POISSON	P_i	P_j	T_{ij}	$C_{ij} + P_j$	99.0	201.5	1.20	-3.6	0.09
46	POISSON	P_i	P_j	S_{ij}	L_{ij}	102.5	196.3	1.45	9.5	0.24
47	POISSON	P_i	P_j	$S_{ij} + P_i$	L_{ij}	101.6	209.8	1.21	1.2	0.10
48	POISSON	P_i	P_j	T_{ij}	L_{ij}	106.2	206.8	1.52	12.0	0.24
49	ZINB	P_i	P_j	S_{ij}	A_{ij}	149.6	1693.4	1.23	50.7	0.10
50	ZINB	P_i	P_j	$S_{ij} + P_i$	A_{ij}	89.8	151.6	1.11	-8.4	0.10
51	ZINB	P_i	P_j	T_{ij}	A_{ij}	168.7	1890.7	1.44	69.9	0.14
52	ZINB	P_i	P_j	T_{ij}	C_{ij}	137.1	1254.9	1.21	37.9	0.10
53	ZINB	P_i	P_j	T_{ij}	$C_{ij} + P_j$	88.1	148.4	1.10	-10.7	0.10
54	ZINB	P_i	P_j	S_{ij}	L_{ij}	145.1	1508.5	1.21	45.7	0.10
55	ZINB	P_i	P_j	$S_{ij} + P_i$	L_{ij}	89.0	158.7	1.10	-9.9	0.10
56	ZINB	P_i	P_j	T_{ij}	L_{ij}	204.7	2946.8	1.31	106.3	0.11
57	ZIP	P_i	P_j	S_{ij}	A_{ij}	106.7	203.4	1.19	16.5	0.15
58	ZIP	P_i	P_j	$S_{ij} + P_i$	A_{ij}	102.2	200.8	1.23	-6.3	0.06
59	ZIP	P_i	P_j	T_{ij}	A_{ij}	116.9	237.3	1.39	23.3	0.17
60	ZIP	P_i	P_j	T_{ij}	C_{ij}	104.8	198.1	1.17	13.9	0.15
61	ZIP	P_i	P_j	T_{ij}	$C_{ij} + P_j$	98.6	189.1	1.20	-10.4	0.07
62	ZIP	P_i	P_j	S_{ij}	L_{ij}	105.0	197.7	1.17	14.6	0.15
63	ZIP	P_i	P_j	$S_{ij} + P_i$	L_{ij}	100.1	193.1	1.19	-6.5	0.07

#	Type	Var 1	Var 2	Var 3	Var 4	MAPE	RMSPE	*Q*	MPE	ME
64	ZIP	P_i	P_j	T_{ij}	L_{ij}	109.8	211.6	1.24	19.4	0.16

Var3 and Va4 are deterrence variables based on spatial population distribution. The best performing combination of variables for each of the four regression model types are highlighted.

The results in Table 5 suggest a clear winner with respect to explanatory variables when estimating models with Census data. Across all four model types, the combination of the distance-weighted Intervening Opportunities T_{ij} and the sum of Competing Migrants plus destination population *(C_{ij} + P_j)* consistently produced top accuracy scores (#37, #45, #53, #61), or any many cases scores that were close to the best. The zero-inflated Negative Binomial version of the model (#53) produced the lowest MAPE and RMSPE in Table 5 while the zero-inflated Poisson model (#58) produced the lowest ME. The OLS models again produced the worst MPE and ME scores.

Table 6. Accuracy of migration model predictions when M ≥ 20 using IRS data.

#	Type	Var 1	Var 2	Var 3	Var 4	MAPE	RMSPE	*Q*	MPE	ME
65	OLS	P_i	P_j	S_{ij}	A_{ij}	60.6	90.8	0.50	24.6	8.43
66	OLS	P_i	P_j	$S_{ij} + P_i$	A_{ij}	56.1	81.4	0.45	21.5	8.27
67	OLS	P_i	P_j	T_{ij}	A_{ij}	62.4	92.2	0.53	25.7	13.69
68	OLS	P_i	P_j	T_{ij}	C_{ij}	60.0	95.1	0.49	24.2	8.28
69	OLS	P_i	P_j	T_{ij}	$C_{ij} + P_j$	55.7	81.0	0.43	21.1	8.13
70	OLS	P_i	P_j	S_{ij}	L_{ij}	59.6	99.6	0.48	24.0	7.73
71	OLS	P_i	P_j	$S_{ij} + P_i$	L_{ij}	55.2	80.2	0.43	20.8	7.58
72	OLS	P_i	P_j	T_{ij}	L_{ij}	60.3	99.3	0.49	24.4	9.49
73	POISSON	P_i	P_j	S_{ij}	A_{ij}	71.7	79.6	5.00	-63.3	0.19
74	POISSON	P_i	P_j	$S_{ij} + P_i$	A_{ij}	71.0	101.2	2.46	-34.1	0.03
75	POISSON	P_i	P_j	T_{ij}	A_{ij}	75.4	84.6	5.71	-61.1	0.21
76	POISSON	P_i	P_j	T_{ij}	C_{ij}	71.9	79.5	4.91	-64.8	0.20
77	POISSON	P_i	P_j	T_{ij}	$C_{ij} + P_j$	69.0	106.1	2.30	-36.5	0.03

#	Type	Var 1	Var 2	Var 3	Var 4	MAPE	RMSPE	*Q*	MPE	ME
78	POISSON	P_i	P_j	S_{ij}	L_{ij}	71.8	79.2	4.90	-64.8	0.20
79	POISSON	P_i	P_j	$S_{ij} + P_i$	L_{ij}	68.8	92.3	2.37	-37.2	0.03
80	POISSON	P_i	P_j	T_{ij}	L_{ij}	73.8	81.4	5.42	-63.9	0.21
81	ZINB	P_i	P_j	S_{ij}	A_{ij}	103.8	285.3	1.84	20.3	0.01
82	ZINB	P_i	P_j	$S_{ij} + P_i$	A_{ij}	63.8	84.3	1.60	-15.8	0.01
83	ZINB	P_i	P_j	T_{ij}	A_{ij}	89.3	167.0	2.68	-8.9	0.04
84	ZINB	P_i	P_j	T_{ij}	C_{ij}	98.9	274.0	1.74	16.1	0.01
85	ZINB	P_i	P_j	T_{ij}	$C_{ij} + P_j$	63.2	85.7	1.51	-16.2	0.01
86	ZINB	P_i	P_j	S_{ij}	L_{ij}	97.5	266.5	1.69	15.8	0.01
87	ZINB	P_i	P_j	$S_{ij} + P_i$	L_{ij}	63.6	87.5	1.48	-14.9	0.00
88	ZINB	P_i	P_j	T_{ij}	L_{ij}	90.6	206.4	2.03	3.2	0.01
89	ZIP	P_i	P_j	S_{ij}	A_{ij}	73.3	104.8	1.75	-5.4	0.01
90	ZIP	P_i	P_j	$S_{ij} + P_i$	A_{ij}	76.7	113.6	2.29	-15.3	0.00
91	ZIP	P_i	P_j	T_{ij}	A_{ij}	81.1	116.4	2.77	-16.4	0.03
92	ZIP	P_i	P_j	T_{ij}	C_{ij}	72.8	106.1	1.65	-5.2	0.01
93	ZIP	P_i	P_j	T_{ij}	$C_{ij} + P_j$	72.5	107.2	2.05	-16.1	0.00
94	ZIP	P_i	P_j	S_{ij}	L_{ij}	72.0	103.7	1.59	-4.4	0.01
95	ZIP	P_i	P_j	$S_{ij} + P_i$	L_{ij}	72.2	102.9	1.99	-15.1	0.00
96	ZIP	P_i	P_j	T_{ij}	L_{ij}	77.4	112.9	2.04	-8.3	0.01

Var3 and Var4 are deterrence variables based on spatial population distribution. Five validation metrics are shown with the best performing combination of variables for each of the four regression model types are highlighted.

In Table 6 values for MAPE, RMSPE, and ME were low relative to Table 4 and Table 5 indicating that IRS migration flows are more readily reproducible using regression models. Interestingly, while the combination of deterrence variables T_{ij} and (S_{ij} + P_j) performed well in Table 6, the combination of S_{ij} + P_i and L_{ij} was the top performer, consistently yielding the lowest error rates across metrics and models (see rows #71, #79, #87, #94). As with the ACS and Census data sources, OLS models estimated using IRS data produced notably poor ME values.

Looking across three data sources and four regression model types, the combination of migration deterrence variables T_{ij} and (S_{ij} + P_j) seem to perform consistently well. Figure 8 and Figure 9 highlight the performance of T_{ij} and (S_{ij} + P_j) relative to other variable combinations for two of the more meaningful and interpretable accuracy metrics: MAPE and ME.

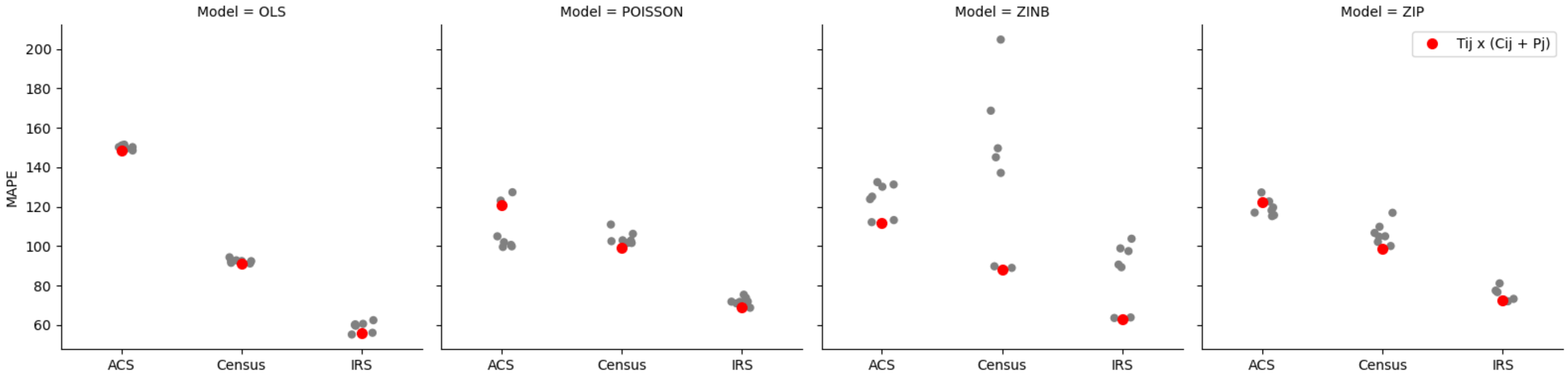

Figure 8. Mean absolute percent error (MAPE) of all models by data source when M > 0 (ACS and Census) and M ≥ 20 (IRS).

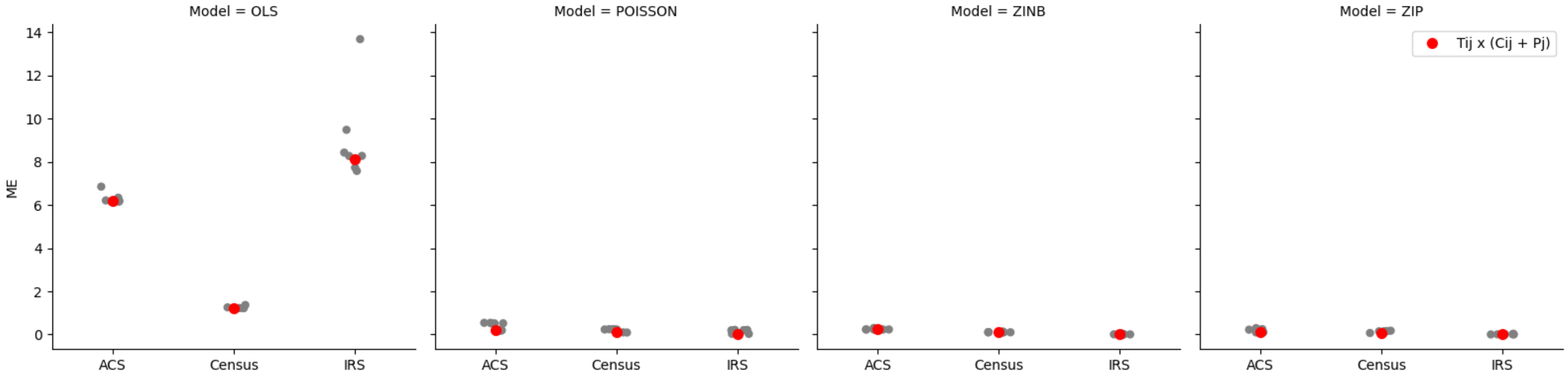

Figure 9. Mean error (ME) of all models by data source when and M = 0.

Selecting the T_{ij} and $(S_{ij} + P_j)$ model from above as the "best of breed" regression model allows for a straightforward comparison to the radiation model. As shown in Figure 10, the radiation model performs quite poorly regardless of data source when validation is constrained to $M \geq 20$. The MLE regression models performed similarly while the OLS approach yielded the best MAPE when using ACS and IRS data. Interestingly, all the regression approaches performed similarly when using the Census migration data.

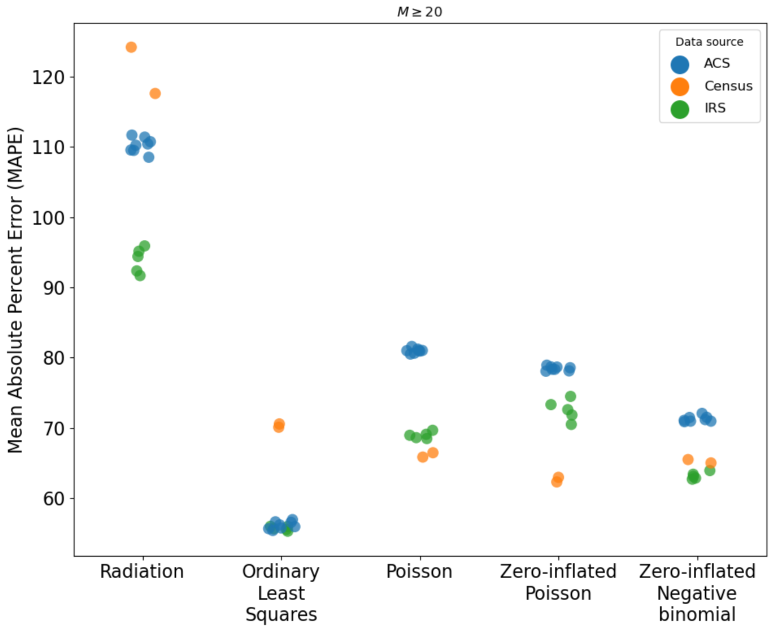

Figure 10. Accuracy of model predictions when $M \geq 20$ using ACS, Census, and IRS migration data.

Figure 11 shows model accuracy when migration flows are greater than zero but less than 20, which excludes the IRS data. First, there is a notable increase in error for all models relative to Figure 10, and a consistent pattern of Census-based models outperforming their ACS counterparts. Another unexpected pattern is the comparable performance of the radiation model, which is markedly different from Figure 10. Finally, the OLS model performs poorly when estimated with ACS data, although all other models perform almost identically when the data source is the same. The zero-inflated Negative binomial model was the most accurate in Figure 11 albeit by a relatively small margin in most cases.

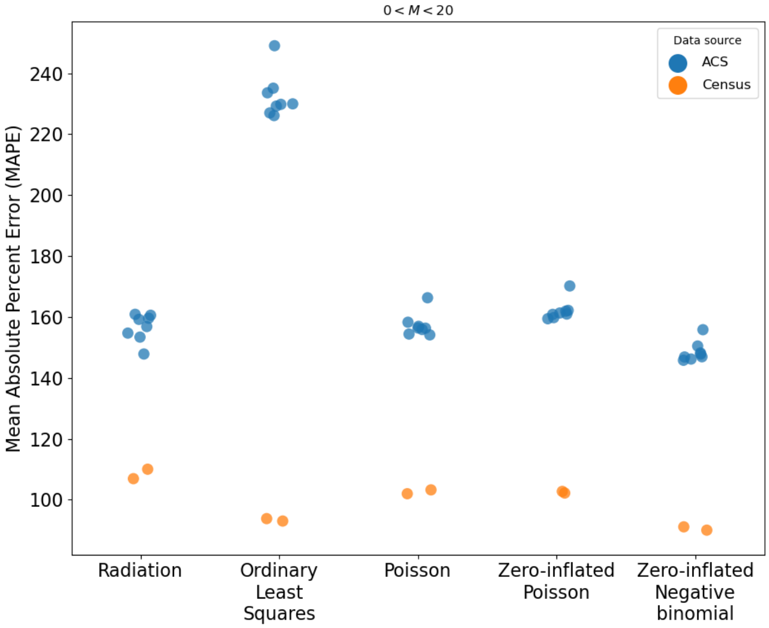

Figure 11. Accuracy of model predictions when 0 < *M* < 20 using ACS, Census, and IRS migration data.

Calculating mean error (ME) when $M = 0$ yields a final set of validation results. Here the radiation model performs quite well, consistently producing the lowest errors regardless of data source. Conversely, the OLS model performs poorly, producing errors an order of magnitude larger than the other models on average. Interestingly, the OLS errors were at least three times lower than when using the Census migration data relative to the ACS and IRS counterparts. The three MLE regression models perform similarly, showing comparable magnitudes of errors which are lowest using IRS data and highest using ACS.

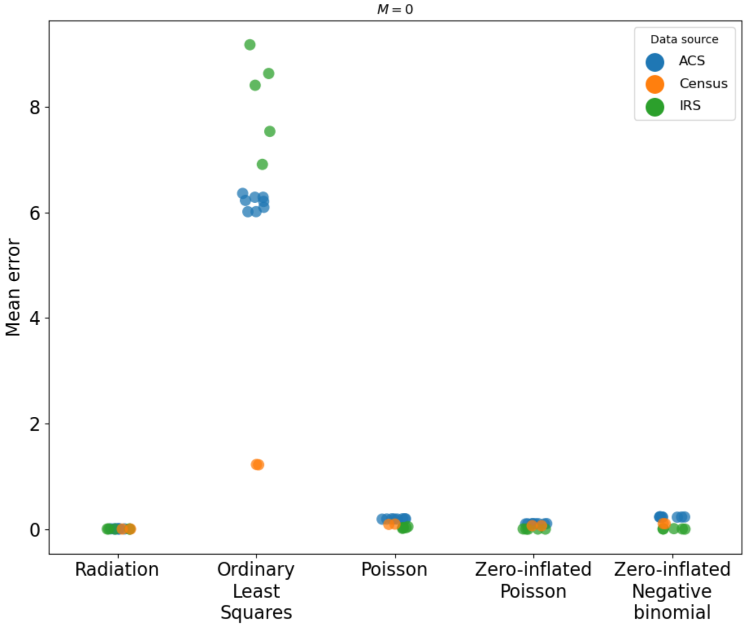

Figure 12. Accuracy of model predictions when $M = 0$ using ACS, Census, and IRS migration data. Regression results shown here use the deterrence variables T_{ij} and $C_{ij} + P_j$.

4.1.3. Discussion

Following on previous work, I find that multiple aspects of forecast accuracy are needed to evaluate alternative model specifications and migration data sources. Assessing prediction errors for $M \geq 20$ alone suggests that the classic OLS spatial interaction model is vastly superior to three MLE approaches. In other circumstances (i.e., cross-sectional analysis) model evaluation might reasonably end with that conclusion. However, evaluating the skill of all four models when observed flows are equal to zero suggests the opposite conclusion. I view this secondary evaluation as essential given that – regardless of data source – more than 90% of all possible county-to-county migration flows are equal to zero each year and the performance of the OLS model must be interpreted in that context.

These results have implications in cases where migration models are needed for long-range projections, and not cross-sectional analyses. When projecting migration forward in time over numerous time steps, small errors accumulate with each iteration of the model, resulting in projections that become more unreasonable over time. The mediocre performance of OLS as shown in Figure 12 may be traced back to an inability to use zero flows in calibration. The natural logarithm of zero is undefined and typically the input data set is truncated to exclude those observations. I followed this convention to preserve consistency with the authors of the radiation model (Simini et al., 2012) who also used a truncated OLS model as a comparator.

My results contradict the conclusions of Simini et al. (2012). I find that the radiation model does poorly at estimating migration flows when $M \geq 20$. The radiation model does seem to be an excellent predictor of zero migration and performs about the same as the regression models when observed migration flows are between zero and 20. Nonetheless, these results show little evidence to support using the radiation model to project migration flows in the U.S.

Additionally, there is evidence that the usual calculations for Stouffer's Intervening Opportunities (S_{ij}) and Competing Migrants (C_{ij}) calculations should be revisited. These results show that two variations of those variables outperform the traditional calculations. For example, holding all other characteristics the same, models that used the variables ($S_{ij} + P_i$) and ($C_{ij} + P_j$) often outperformed models that included S_{ij} and C_{ij}, respectively. In the case of Intervening Opportunities, this could reflect population at the origin providing some amenity that discourages potential migrants from leaving. Conversely, population at the destination should themselves out-compete potential migrants for employment or housing given the disparity in proximity and access to information. Moreover, the distance-weighted versions of S_{ij} and C_{ij} – T_{ij} and L_{ij} respectively – were the *best* performing variants of Intervening Opportunities and Competing Migrants. Although the patterns were not entirely consistent across model types and data sources, there seem to be both empirical and theoretical justifications for inclusion of variants of Stouffer's original ideas.

4.2. <u>Homophily in migration</u>

4.2.1. Race

4.2.1.1. Methods

I first parse migration data by race and report correlation coefficients that demonstrate homophily in migration destination choices. Next, I apply the Intervening Opportunities-Competing Migrant model (IO-CM) identified in section 4.1 to estimate migration flows by race: I perform model estimation in this section using the 1990 Census migration data based primarily on the results shown previously in Figure 11, which show that models estimated using Census data better replicate observations than models estimated using ACS data. In addition, the 1990 Census data are fully cross tabulated by race and age, which permits further sub-setting of migration flows by age in Section 4.2.2. Equation 14 describes the modeling approached used to estimate race-specific migration flows in this section:

$$M_{ij,r,d,t+1} = \frac{P_{i,r,t}^{\beta_1} \cdot P_{j,r,t}^{\beta_2}}{T_{ij,r,t}^{\beta_3} \cdot \left(C_{ij,r,t} + P_{ij,r,t}\right)^{\beta_4}}$$

Equation 14. The Intervening Opportunities-Competing Migrants (IO-CM) spatial interaction model.

where $M_{ij,r,t+1}$ is the estimated number of migrants between the origin *i* and destination *j* of race *r* between year *t* and *t+1*; $P_{i,r,t}$ and $P_{j,r,t}$ are the population of race *r* at the origin and destination, respectively, in year *t*; $T_{ij,r,t}$ is the distance-weighted Intervening Opportunities calculation from Equation 7 considering only race *r* for year *t*; and $\left(C_{ij,r,t} + P_{ij,r,t}\right)$ is a modified Competing Migrants calculation (see Table 3) considering only race *r* for year *t*.

4.2.1.2. Results

The Spearman correlation coefficients relating migration into a location and the population at that location by race and ethnicity are presented in Table 7: Across all migration data sets the within-race correlation between incoming migration and population was greater than the correlation between the same migration flows and *total* population. The correlations shown in Table 7 suggest that predictive capability may be improved by articulating migration flows by race and then summing the resulting predictions. An additional benefit of partitioning migration in this manner is that the racial characteristics of gross migration flows are retained, which is relevant for policy making that considers environmental justice (Bowen et al., 1995; Bowen & Haynes, 2000).

Table 7. Correlation by race (*r*) between incoming migration and population.

		Spearman correlation with M_r	
Migration data source	r	$P_{j,r}$	P_j
1990 Census			
	White	**0.948 (0.945, 0.952)**	0.936 (0.932, 0.940)
	Black	**0.948 (0.945, 0.952)**	0.733 (0.716, 0.749)
	Other	**0.888 (0.881, 0.896)**	0.722 (0.705, 0.739)
	M_j		0.948 (0.944, 0.951)
2000 Census			
	White	**0.961 (0.958, 0.964)**	0.946 (0.942, 0.950)
	Black	**0.956 (0.953, 0.959)**	0.758 (0.743, 0.773)
	AIAN	**0.899 (0.892, 0.905)**	0.689 (0.670, 0.708)
	API	**0.891 (0.883, 0.898)**	0.859 (0.850, 0.868)
	M_j		0.960 (0.957, 0.963)
2011-2015 ACS			
	White Alone	**0.939 (0.935, 0.943)**	0.929 (0.924, 0.934)
	Black Alone	**0.884 (0.876, 0.892)**	0.769 (0.754, 0.783)
	Asian Alone	**0.769 (0.754, 0.783)**	0.729 (0.713, 0.746)
	Other	**0.839 (0.828, 0.849)**	0.785 (0.771, 0.798)
	M_j		0.936 (0.931, 0.940)
	Hispanic	**0.836 (0.825, 0.846)**	0.756 (0.741, 0.771)
	Not Hispanic	**0.939 (0.934, 0.943)**	0.933 (0.928, 0.937)
	M_j		0.936 (0.932, 0.941)

The 95 percent confidence intervals are shown in parentheses. $P_j$is the total population at the destination while $P_{j,r}$ is the total population of race *r* only. $M_j$is the total incoming migration (all races) at the destination.

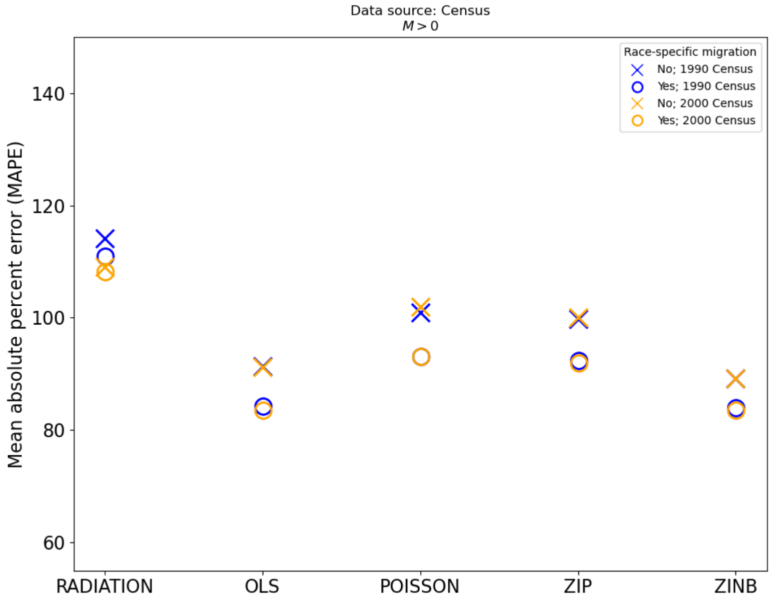

Figure 13. Comparison of IO-CM accuracy with and without race-specific migration using Census data when $M > 0$.

Figure 13 shows that total non-zero migration flows predicted by Census-calibrated models were generally more accurate when migration for each race was calculated independently. Interestingly, Figure 14 shows that predictions of zero flows were largely unchanged, with the exception the OLS model error which roughly doubled when migration was articulated by race.

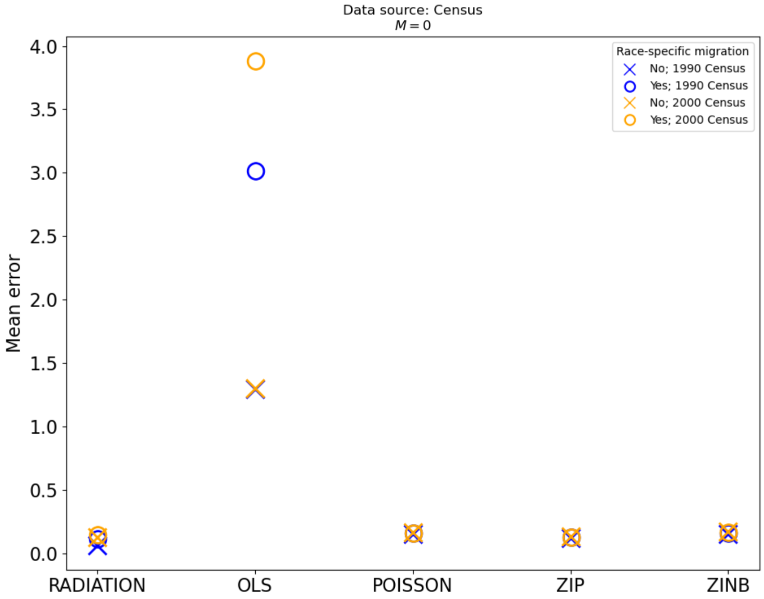

Figure 14. Comparison of IO-CM accuracy with and without race-specific migration using Census data when *M* = 0.

4.2.1.3. Discussion

Total migration is better predicted by modeling race-specific migration flows, and then summing the results. I propose that this result reflects the influence of homophily on migration decisions. As the population of a race increases at a location, the more socially connected that location becomes for other members of that race. Information about jobs and housing opportunities may be acquired through social networks (DaVanzo, 1981), which in turn are shaped largely by race (McPherson et al., 2001; Mele, 2021). A desire to leverage social capital at the individual level may also influence migration decisions. This idea does not preclude other motives as the primary migration driver, but rather considers that a job hunt, for example, may be limited to places where the job seeker has preexisting social network connections.

The role of race in migration decisions among African-Americans has been reported in much earlier work (Cebula et al., 1973; McHugh, 1988; Pack, 1973). Frey and Liaw (2005) found race-ethnicity populations at destinations to be a significant factor in migration flows for Hispanic, Black and Asian population groups. Race is infrequently incorporated into migration studies or population projections.

4.2.2. Age homophily, urban preference and labor markets

4.2.2.1. Methods

The results of section 4.2.1 demonstrate a meaningful link between racial homophily and migration. In this section I further explore this connection by introducing an additional variable P_j^* which represents the same-race population of the destination labor market. I include this variable to reflect the that the likelihood, number, or strength of social network connections not only for the destination county but also for surrounding area would function as a "pull" force for potential migrants.

I also include two additional dummy variables that capture important aspects of migration found in the literature. The first indicates whether an origin-destination pair would result in an *intra*-labor market move or not. Scholars have long noted a general dichotomy between shorter, more frequent "adjustment" moves with the goal of acquiring new housing or access to amenities and longer, less frequent moves which result from changing jobs, attending college, or other significant life course events (e.g., Clark & Onaka, 1983; Plane et al., 2005). I expect the coefficient for intra-labor market moves to be positive and statistically significant.

The second dummy variable indicates whether the destination county is urban or not. I use the U.S. Census Metropolitan Statistical Areas to define urban counties since that definition is rigorously derived and widely used. This dummy variable is needed to capture the secular trend of urbanization in the United States (Cromartie, 2020; D. McGranahan et al., 2010), particularly for early career, young adults (Plane et al., 2005; Plane & Heins, 2003; Plane & Jurjevich, 2009).

The use of both the 1990 Census and 2011-2015 ACS migration data to estimate homophily based migration models requires additional methodological considerations. The 1990 Census data include migration flows crossed by race (n = 3) and age group (n = 17), so unique models can be estimated for each race/age combination. The implementation of racial homophily is straightforward: a simple binning of migrant and populations by race. To affect migration with age homophily, destination populations (Pj and Pj*) were weighted using age group-to-age group correlations for each age group migration model. The values used to weight destination populations are visualized in Figure 15, i.e., a given row represents the set of weights used to calculate destination populations for the corresponding age-group on the vertical axis.

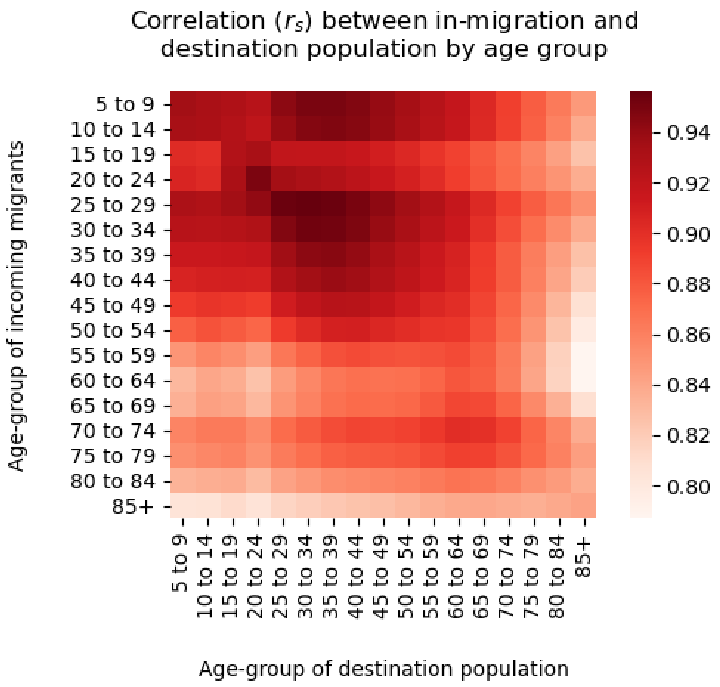

Figure 15. Correlation (*rs*) between in-migration and destination population by age group in 1990 Census migration data.

The final homophily model of migration using the 1990 Census data is given as:

$$M_{ij,r,\mathrm{a},t+1} = \frac{P_{i,r,\mathrm{a},t}^{\beta_1} \cdot P_{j,r,\mathrm{w},t}^{\beta_2}}{T_{ij,r,\mathrm{w},t}^{\beta_3} \cdot \left(C_{ij,r,\mathrm{w},t} + P_{j,r,\mathrm{w},t}\right)^{\beta_4}} \cdot URB_j^{\beta_5} \cdot LM_{ij}^{\beta_6} \cdot P_{j,r,w}^{*\beta_7}$$

Equation 15. The homophily model of migration using 1990 Census migration data.

where $\widehat{M}_{ij,r,a,t+1}$ is the estimated number of migrants between the origin *i* and destination *j* belonging to a given race *r* and age group *a* between years *t* and *t+1*; $P_{i,r,a,t}$ is the population at *i* belonging to race *r* and age group *a* for year *t*; $P_{j,r,w,t}$ is the age-weighted population *w* at *j* belonging to race *r* for year *t*; $T_{ij,r,w,t}$ is the distance-weighted

Intervening Opportunities calculation from Equation 7 considering only the age-weighted population w of race r for year t; $\left(C_{ij,r,w,t} + P_{j,r,w,t}\right)$ is a modified Competing Migrants calculation (see Table 3) considering only the age-weighted population w of race r for year t; URB_j is binary variable indicating whether the destination county is urban; LM_{ij} is a binary variable indicating whether the destination county j is contained within the same labor market as the origin county i; and $P^{*}_{j,r,w}$ is the age-weighted population belonging to race r that resides outside of the destination county j but within the same labor market as j.

4.2.2.2. Results

I evaluated the impact of age-homophily as well as the additional variables LM_{ij}, URB_j, and $P^{*}_{j,r,w}$ by incrementally comparing the mean absolute percentage error (MAPE) of those model predictions against the predictions of the ZINB race homophily model show in Figure 13. As shown in Table 8, the additional parsing of migration by age group increases the MAPE by nearly 15%. The additional variables nominally reduce the overall MAPE, however the ZINB race-only homophily model from Figure 13 remains the best performer. The results suggest that role of age homophily is not well described by the variables in included here, or that age homophily is not a significant factor in migration decisions.

Table 8. Incremental contribution of additional model variables.

Equation	**Mean absolute percentage error (MAPE)**
$\boldsymbol{P_{i,r} \cdot P_{j,r} \cdot T_{ij,r} \cdot \left(C_{ij,r} + P_{j,r}\right)}$	*83.8*[†]
$\boldsymbol{P_{i,r,a} \cdot P_{j,r,w} \cdot T_{ij,r,w} \cdot \left(C_{ij,r,w} + P_{j,r,w}\right)}$	98.5
$\boldsymbol{P_{i,r,a} \cdot P_{j,r,w} \cdot T_{ij,r,w} \cdot \left(C_{ij,r,w} + P_{j,r,w}\right) \cdot LM_j}$	97.65
$\boldsymbol{P_{i,r,a} \cdot P_{j,r,w} \cdot T_{ij,r,w} \cdot \left(C_{ij,r,w} + P_{j,r,w}\right) \cdot LM_{ij} \cdot URB_j}$	97.08
$\boldsymbol{P_{i,r,a} \cdot P_{j,r,w} \cdot T_{ij,r,w} \cdot \left(C_{ij,r,w} + P_{j,r,w}\right) \cdot LM_{ij} \cdot URB_j \cdot P^*_{j,r,w}}$	96.9

[†] This value is from the ZINB results in Figure 13.

Model estimation in this chapter results in unique coefficients for both the *zero-inflation* and *count* parts of the ZINB migration model. The zero-inflation component is estimating the probability that a migration flow between an origin and destination will be equal to zero, i.e., $\Pr(\widehat{M}_{ij} = 0)$. The count component is estimating the expected number of migrants between two locations *given* the probability that the flow is zero, i.e., $\widehat{M}_{ij} \mid Pr\left(M_{ij} = 0\right)$. Coefficients for T_{ij} and $S_{ij} + P_i$ are shown below. The remaining coefficients are included in the Appendix.

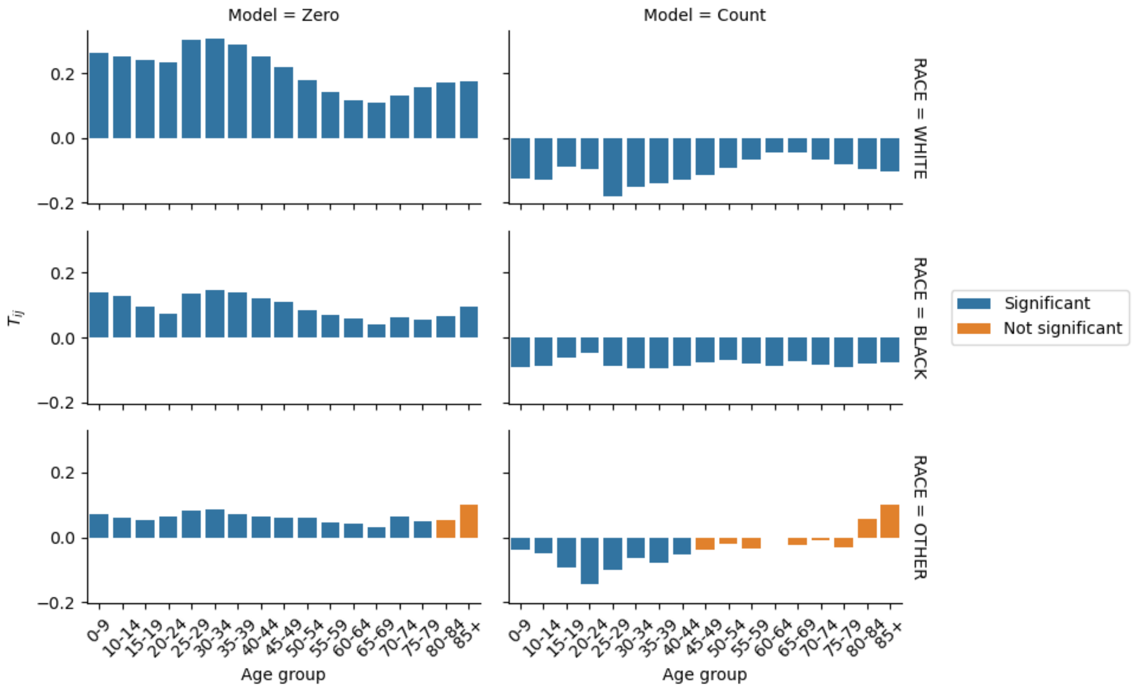

Figure 16. Coefficients of Tij by race and age group for the zero-inflation and count components of the ZINB model.

For the zero-inflation component of the model, the sign of the coefficients for T_{ij} are positive for all age and race groups, which is expected given the conceptual basis for this variable. All coefficients were statistically significant ($\alpha = 0.5$), except for the two oldest age groups (80-84 and 85+) in the OTHER race group. For the variable T_{ij} in the count portion of the model, the sign of all statistically significant coefficients ($\alpha = 0.5$) is negative, which is again consistent with the conceptual basis of the variable. In the OTHER race group, the variable T_{ij} was not statistically significant for all age groups greater than 40-44. For the two oldest age groups in the OTHER race category, the sign of the coefficient was positive.

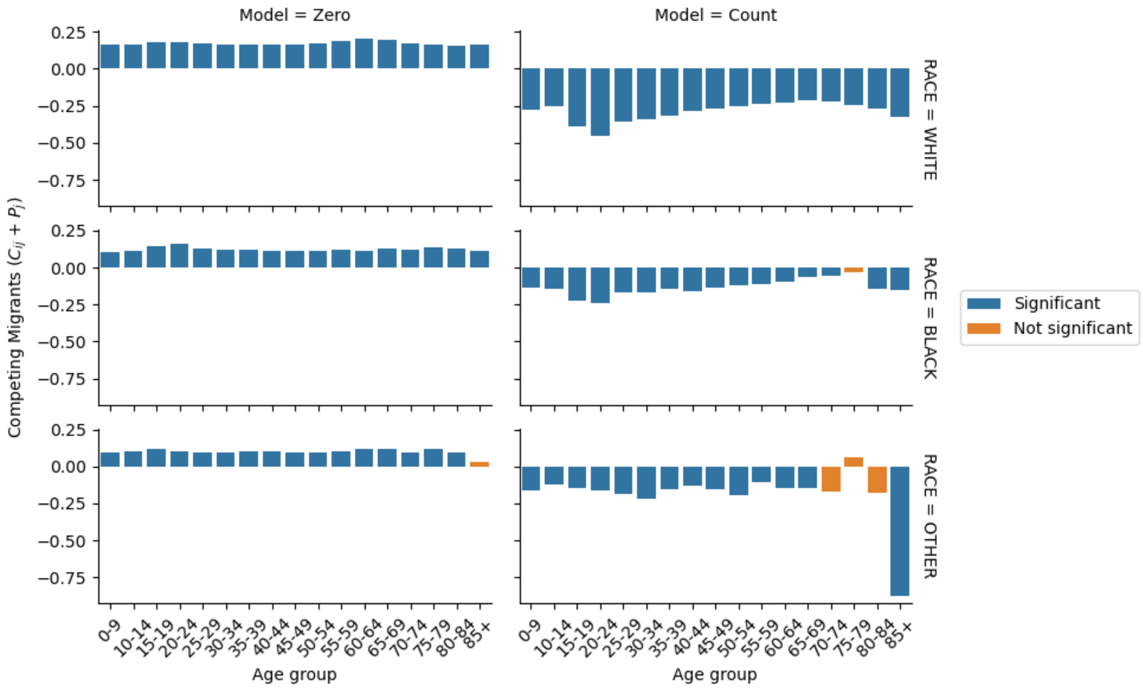

Figure 17. Coefficients of $(C_{ij} + P_j)$ by race and age group for the zero-inflation and count components of the ZINB model.

Coefficients of variable $C_{ij} + P_j$ in the zero-inflation component of the model were positive for all combinations of race and age group, indicating that, as expected, the probably of $\widehat{M}_{ij} = 0$ increases as the number of intervening opportunities increases.

For the count component of the ZINB model, nearly all coefficients for the variable $C_{ij} + P_j$ statistically significant ($\alpha = 0.5$) and negative. The coefficient for the 75-79 age group in the OTHER race category was the only coefficient that was positive. That same combination was not statistically significant, either, joining the 70-74 and 80-84 age groups in the OTHER race category, as well as the 75-79 age group in the BLACK race group.

4.2.2.3. Discussion

The degraded performance of the final age-race homophily migration model (Equation 15) relative to the race only homophily model was surprising. Numerous correlations between migrating and destination age groups were found (Figure 15) and suggested an empirical relationship that was consistent with the homophily principle. It's possible that unlike race, the age group correlations reflect migration outcomes driven by other factors. From an analytic perspective, the reliance on age homophily and the additional variables shown in Table 8 have little support. However, articulating migration flows by age and race is critical in the context of population projections and moving forward with Equation 15 is therefore justified, particularly since the literature provides no examples of what constitutes a "good" migration model in that context.

The additional results in section 4.2.2.2 largely conformed to expectations. The novel deterrence variables were largely statistically significant and of the correct sign (Figure 16 and Figure 17). There was a somewhat consistent pattern of the coefficients for those variables being statistically insignificant and of the unexpected sign. For example, in Figure 16 the sign of the COUNT coefficient for T_{ij} is positive for the two oldest age groups in the OTHER row. This is counterintuitive since we would expect T_{ij} to attenuate migration as its value increases, i.e., I expect the sign to be negative. Of course, in that case the coefficient was not statistically significant and not particularly concerning. However, the results do suggest a lack of explanatory power for the older age groups overall. Previous work has found that older American and retirees frequently move to more rural areas rich with natural amenities (D. McGranahan, 1999; D. A.

McGranahan, 2008) and the results presented here may be improved by including corresponding variables.

Insignificant terms in the OTHER race group likely reflect that this label is really an aggregate of multiple race groups, e.g., Asian and Native American. The ACS migration data include an Asian breakout; however, it does not provide the race and age cross tabulation needed to reproduce the analyses in this section. This is an area of research that may benefit greatly from synthetic migration data (e.g., Granberry et al., 2018) which would resolve the major limitations of the ACS (no cross tabulations), Census (older vintage) and IRS (no demographic information) data sources.

4.3. Comparing sub-national population projection

4.3.1. Methods

Using the homophily based migration model from section 4.2.2 I produced population projections for all U.S. counties through the year 2020. At each annual time step births were calculated by multiplying projected percent changes in fertility (Wittgenstein Centre for Demography and Global Human Capital, 2018a) by historical birth rates (Centers for Disease Control and Prevention, 2020b). Annual deaths were similarly calculated by multiplying projected percent changes in mortality (Wittgenstein Centre for Demography and Global Human Capital, 2018b) by historical death rates (Centers for Disease Control and Prevention, 2020a). I calculate and compare the mean absolute percentage errors (MAPE) for that and several other population projections using the estimated 2020 county population from U.S. Census Bureau. Population projections were selected using the simple criteria that they were resolved at the state or county level for the entire conterminous United States. I assessed model performance objectively at the state and county level using typical MAPE values described by Smith et al (2013). A summary of the models assessed in this dissertation is provided in Table 9 (Appendix).

4.3.2. Results

State-level MAPE values for 11 population projections are shown in Figure 18. As expected, the accuracy of a given population projection (or a given scenario) is largely a function of the launch year. Smith et al. (2013) characterize a typical MAPE value for state-level population projections as 6% per decade of the projection horizon. Surprisingly, virtually all projections were better than the 6% threshold and 4% per decade appeared to be a better central estimate of MAPE values. None of the projections were better than 2% MAPE per decade.

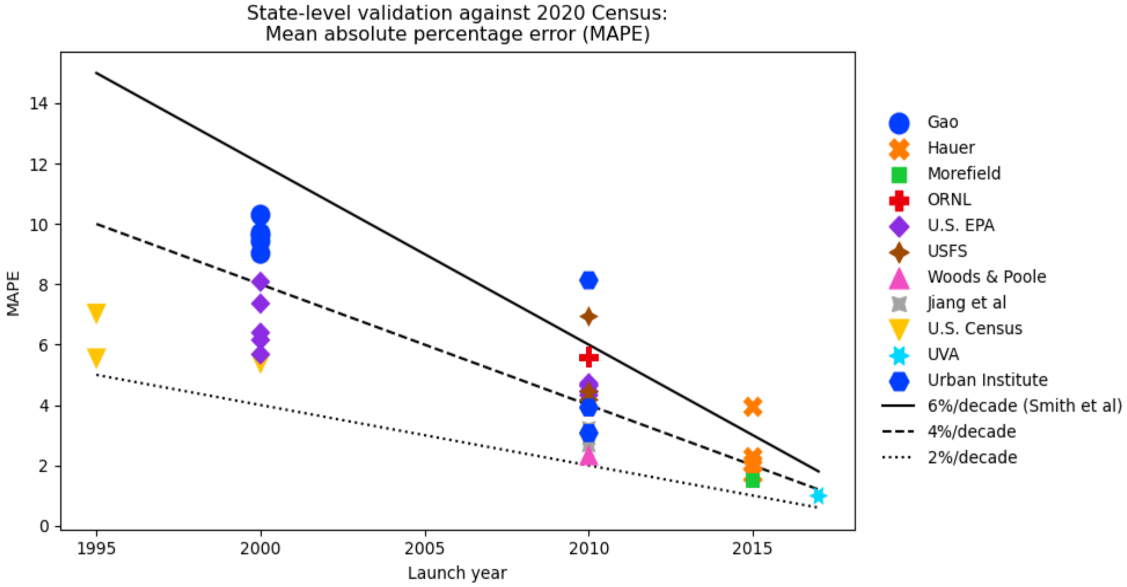

Figure 18. Assessing state-level projection errors using MAPE.

County level MAPES are shown in Figure 19 and displayed similar characteristics to the state-level results. Smith et al. suggested 12% per decade as a typical MAPE value for county population projections, however my results show that 8% per decade are more representative of typical MAPE values at the county level. At this geographic scale, none of the projections were better than 4% MAPE per decade. Like Smith et al, tightening the scale to county level leads to MAPE values approximately double those at the state level.

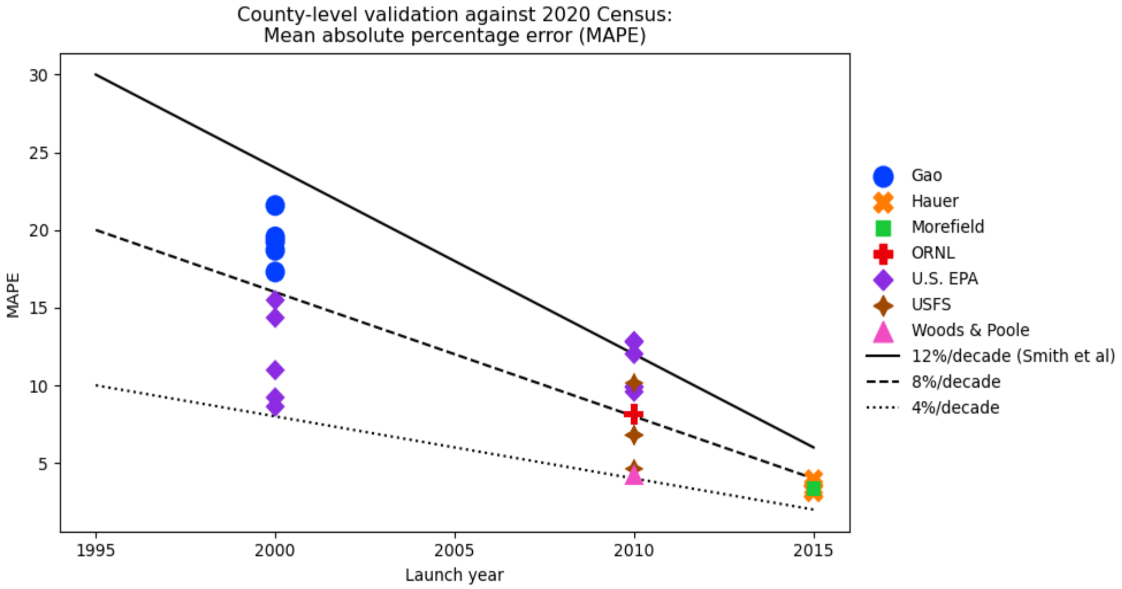

Figure 19. Assessing county level projection errors using MAPE.

The spatial pattern of errors in my final demographic model reveals stark patterns in the distribution of errors, with considerable variance in sign and magnitude over short distances. While projection error was ±2% for many counties – which corresponds to the 4% per decade threshold highlighted previously – large groups of counties were beyond the 12% threshold, which is relatively poor. Total population was substantially over-projected (shown in blue) for several counties spanning western North Dakota and eastern Montana. This contrasts with much of the western United States, where populations were frequently *under* predicted (shown in red). Much of the eastern 1/3rd of the United States lacked any consistent pattern beyond most counties falling into the ±8% error per decade groupings.

The results show in Figure 20 also highlight persistent challenge in producing accurate population projections. For example, the systematic overprediction of population in western North Dakota and eastern Montana appears to be the result of excess migration predicted by my model. The recent boom in fracking activity in that region has resulted in a surge of predominantly male laborers. My model does not take this unique economic context into account and projects the migration of males and females to this region with equal likelihood.

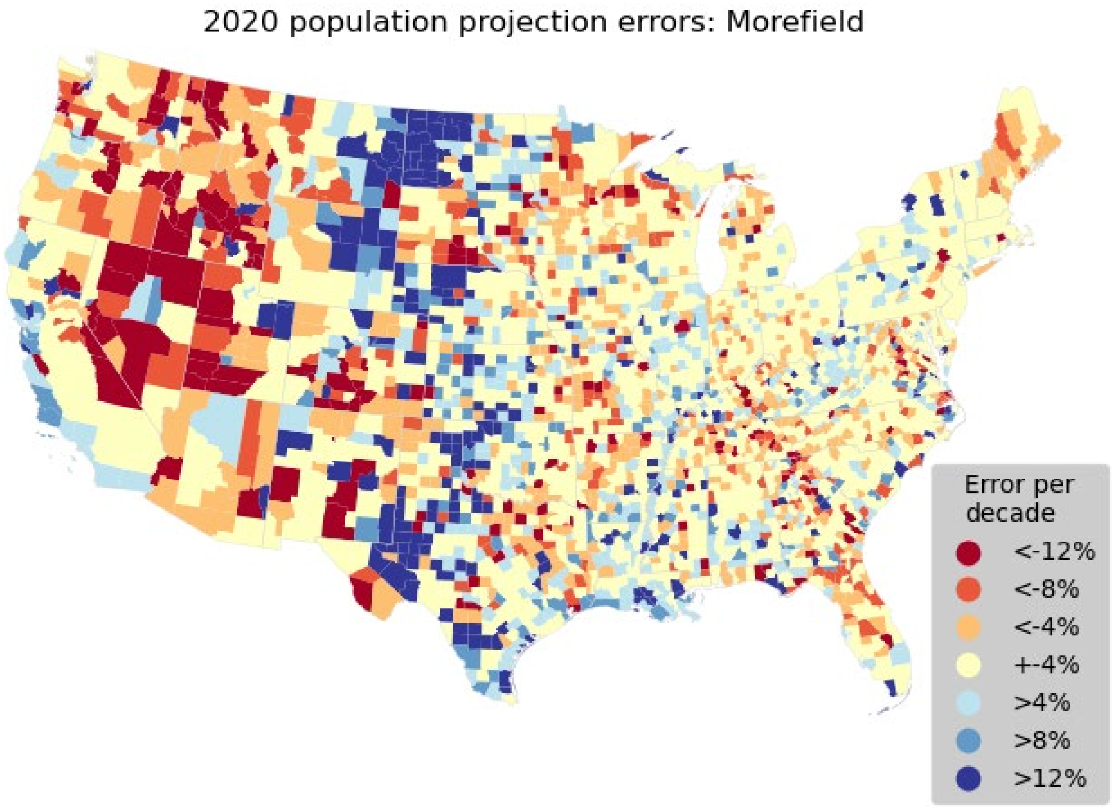

Figure 20. County level projection errors of the model described in this dissertation.

4.3.3. Discussion

The results of this section show that migration can be modeled as a predominantly social process and – in the context of a complete demographic model, at least – produce reasonably accurate results. Using a homophily model of migration, I was able to produce population projections with five-year MAPE values of 1.49% and 3.33% at the state and county level, respectively. Those errors are competitive with the population scenarios produced by Hauer (2019) and, somewhat surprisingly, are characterized by a very similar spatial distribution (see Figure 25 in Appendix). The model produced in this dissertation is also competitive across all other projections of similar spatial and temporal scopes.

I also find that the ‘typical’ MAPE values suggested by Smith et al. (2013) are misleading in the context of the more contemporary projections assessed in this dissertation. In fact, their per-decade error rates of 6% (state) and 12% (county) better characterize the *least* accurate projections currently available validated against 2020 population. Respective error rates of 4% and 8% for state and county projections seem to better capture typical expectations. My results also bring into focus a potential definition of ‘best available’ projections. Based on Figure 18 and Figure 19, MAPE rates of 4% and 2% per-decade represent aspirational targets for state and county level projections, respectively.

5. CONCLUSION

This dissertation advances migration research by demonstrating the viability of a new modeling approach focused on mobility as a social process, instead of an economic one. Additional research contributions are made in the field of population geography, first by demonstrating that credible population projections can be produced using this new theoretical framework for migration. The second such contribution is an assessment of the accuracy of population projections, establishing clear benchmarks for evaluating past and future projection efforts.

Population projections have become an essential component of policy relevant environmental and global change research. Understanding the potential impacts of large-scale environmental hazards such as climate change requires insights into the geography of population: the location and characteristics of the U.S. population over time. My research sought to improve understanding of migration in the context of population projections by addressing three research questions.

The first research question evaluated new spatial variables that more effectively capture migration deterrence as a social process rather than simple Euclidean distance. Included in this evaluation was an original approach to modeling human mobility, the radiation model, and the comparisons were made against three different migration data sources. My assessment of these various methods in the context of different data sources proved useful, showing a diverse range of model performance depending on model form, variable choice, and the data source used to estimate the model. I was also able to show that the radiation model performs poorly relative to more traditional regression approaches, but that even those frequently cited methods can still benefit from new twists such as zero-inflation. I was able to demonstrate that a new model – based on classic ideas of Stouffer and leveraging relatively untested zero-inflation statistical modeling – shows tremendous promise as a starting point for future migration research efforts.

My second research question explored the idea of modeling migration as a social process rather than an economic choice. Using the homophily principle as a theoretical framework did indeed improve model performance, lending credence to the notion that people make migration choices using information outside of strict financial calculations (as has been suggested by others). Translating this idea into model mechanics is trivial: simply partitioning migration calculations by race. Doing so provided a measurable boost to model performance with the added benefit of a self-accounting of population by race; a desirable characteristic unto itself. This line of my research was carried a step further, integrating age homophily into the final model as well as well-documented preferences for urban areas for certain age groups. The final migration model includes predictors that are known influences of migration behavior but can accommodate additional variables of interest over longer time scales (unlike fixed-rate models). For example, the impact of climate change on population is a concern at multiple levels of governance and may be manifested though multiple pathways. Increasing frequency of wildfires may generate pulses of out-migration from affected areas. Hurricanes similarly have resulted in mass migration events from coastal cities such as New Orleans. And the gradual pressure of sea-level rise seems likely to drive population inward, away from the most low-lying coastal areas of the United States in the coming decades. The timing and scope of the effect of each of these drivers is an important area of research, and each may be represented by the migration model described in this dissertation by including exogenous data or model outputs as additional predictor variables..

The last portion of my research integrates this final model into a full demographic model and produces projections that let me perform a final assessment. My results show clearly that this new modeling approach shows enormous potential and can be used to produce demographically detailed population projections that perform well compared to similar efforts. Moreover, my assessment of model accuracy reveals useful MAPE benchmarks for evaluating population projection accuracy and utility going forward.

The modeling approach demonstrated in this dissertation forgoes economic variables shown to influence migration patterns for two reasons. First, this is simply an acknowledgment that there are no county level projections of those variables frequently found in the literature such as wages and housing prices. Second, this dissertation sought to demonstrate the utility of a relatively simple migration model based on singular sociological principle. The methodological choices in this work should not be construed as disregard for the vast economic migration literature, but rather a renewed focus on the phenomenon of migration through a lens of human geography.

An important limitation of the work presented here is the inherent uncertainty of population projections. The out-of-sample accuracy assessments themselves are constrained by data availability with some components of change going back only to 1995. The assessments presented here provide a measure of accuracy that may be presently useful for model selection, but subsequent evaluations may support different conclusions regarding the suitability of a given model or scenario. Sub-national population trajectories are subject to sudden social, economic, and political forces and are thusly difficult to predict with any certainty. A rapidly changing climate may affect the habitability of arid regions of the country, for example, by constraining water availability. Similarly, sea-level rise may dramatically impact housing choices along large swaths of the U.S. coastline. While these outcomes are virtually impossible to predict, a modeling framework such as the one presented here can facilitate assessments of those environment-migration interactions through scenario analysis.

The potential impacts of those large-scale environmental disruptions on domestic migration stand out as an area of need for future research. Insights into the timing and magnitude of sea-level rise impacts on migration choices, for example, are relatively thin considering the potential scope of that phenomenon. However, I would argue that a more immediate need with respect to future scholarship is the development of a unified migration data source. This next step should be easily achievable with existing methods and tools already used to develop synthetic population. A thematically and spatially detailed migration data source that combines information and attributes from multiple resources would improve comparability, transparency, and reproducibility in the field.

APPENDIX

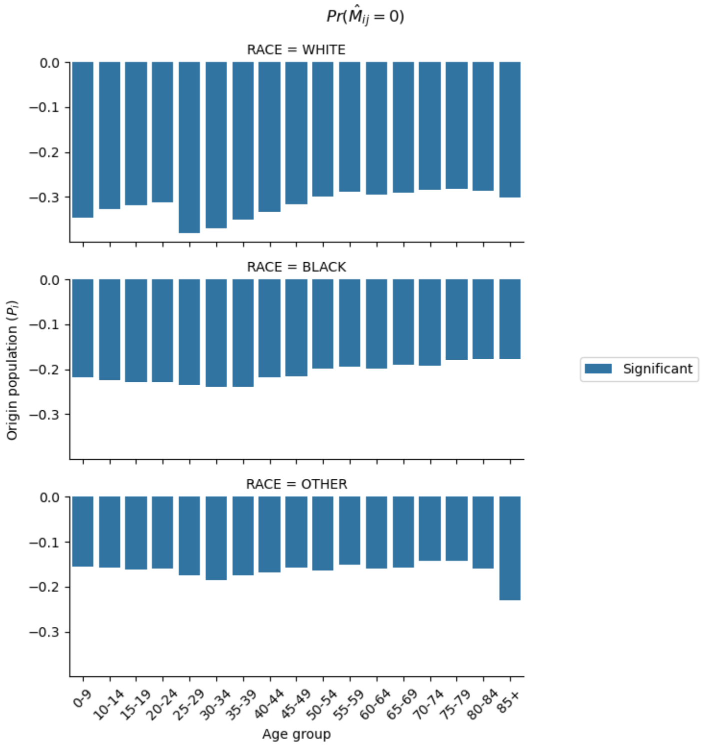

Figure 21. Coefficients of P_i by race and age group for the zero-inflation component of the ZINB model.

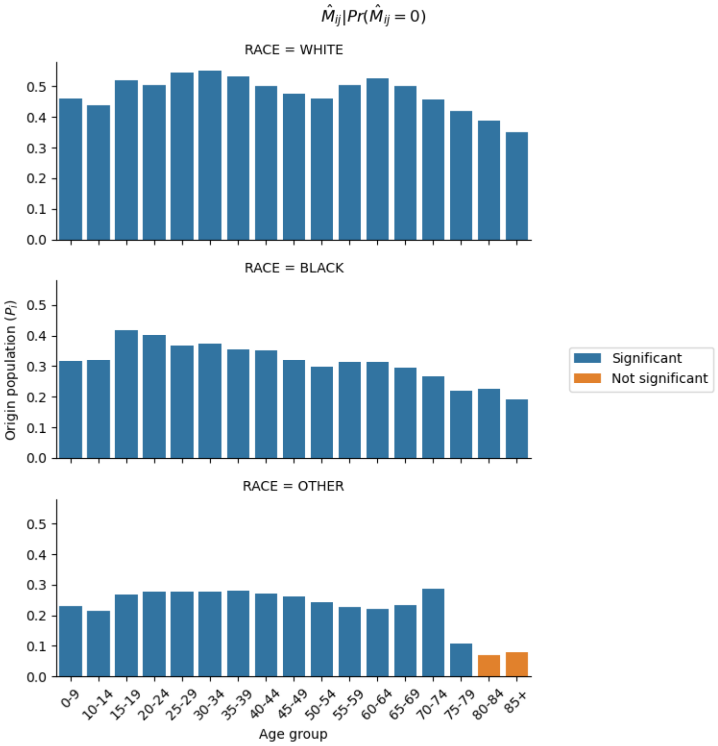

Figure 22. Coefficients of P_i by race and age group for the count component of the ZINB model.

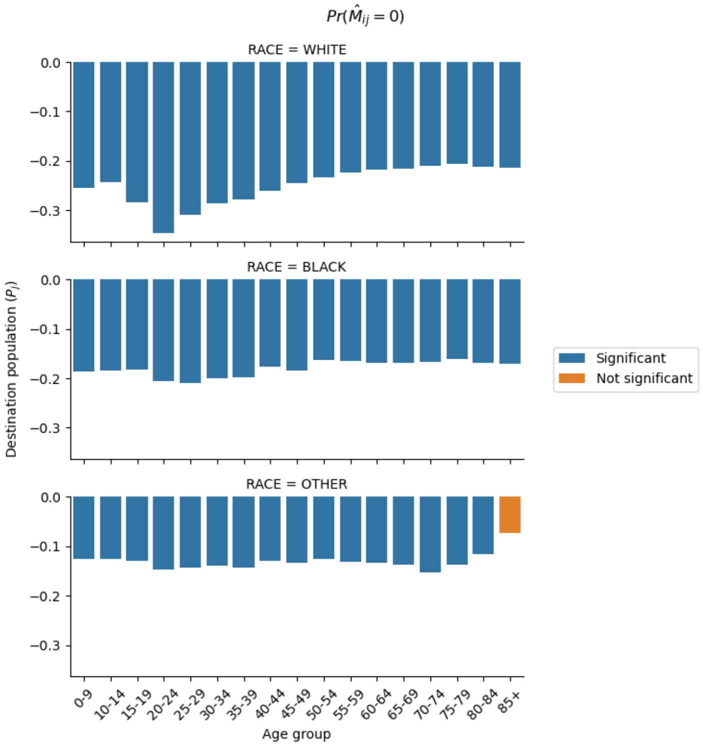

Figure 23. Coefficients of P_j by race and age group for the zero-inflation component of the ZINB model.

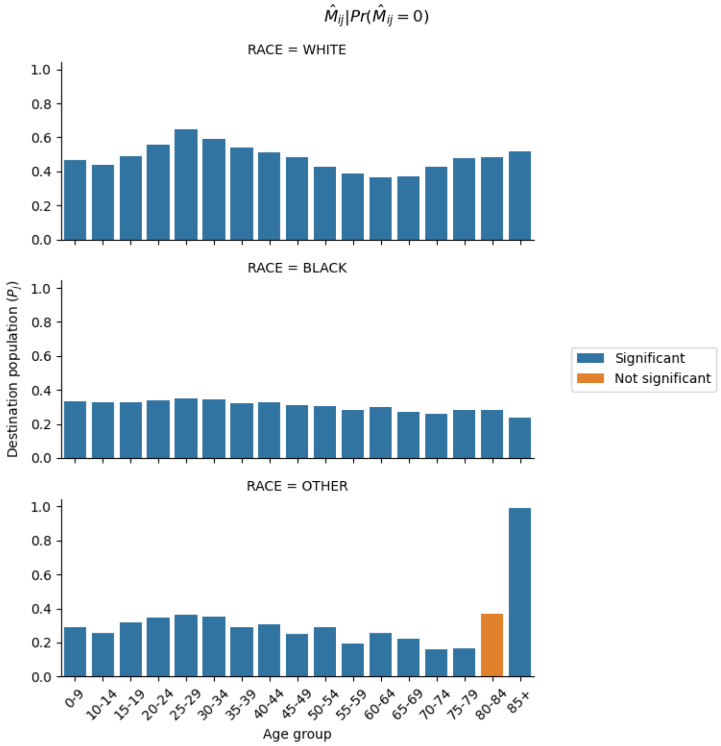

Figure 24. Coefficients of P_j by race and age group for the count component of the ZINB model.

Table 9. Summary of population projections assessed in this dissertation.

Name	Year published	# of projections	Spatial resolution	Temporal attributes	Thematic attributes	Launch year
U.S. EPA	2009	6	Counties	Decadal through 2100	Total population	2000
U.S. EPA	2017	6	CBSAs and counties	Decadal through 2100	Total population	2010
Gao	2017	5	1 km	Decadal through 2100	Urban and rural	2000
Mckee et al.	2015	1	30 arc-second	2030 and 2050 only	Total population	2010
University of Virginia	2017	1	States	Decadal through 2040	Age and race	2017
Urban Institute	2017	27	Commuting zones	Semi-decadal through 2040	Age and race	2010
USFS	2019	5	Counties	Semi-decadal through 2070	Per Capita Income	2010
Jiang et al.	2020	5	States	Decadal through 2100	Urban and rural	2010
Hauer	2019	5	Counties	Semi-decadal through 2100	Age, sex, and race	2010
U.S. Census Bureau	1997	2	States	Semi-decadal through 2025	Total population	1995
U.S. Census Bureau	2005	1	States	Annual through 2030	Age & gender	2004

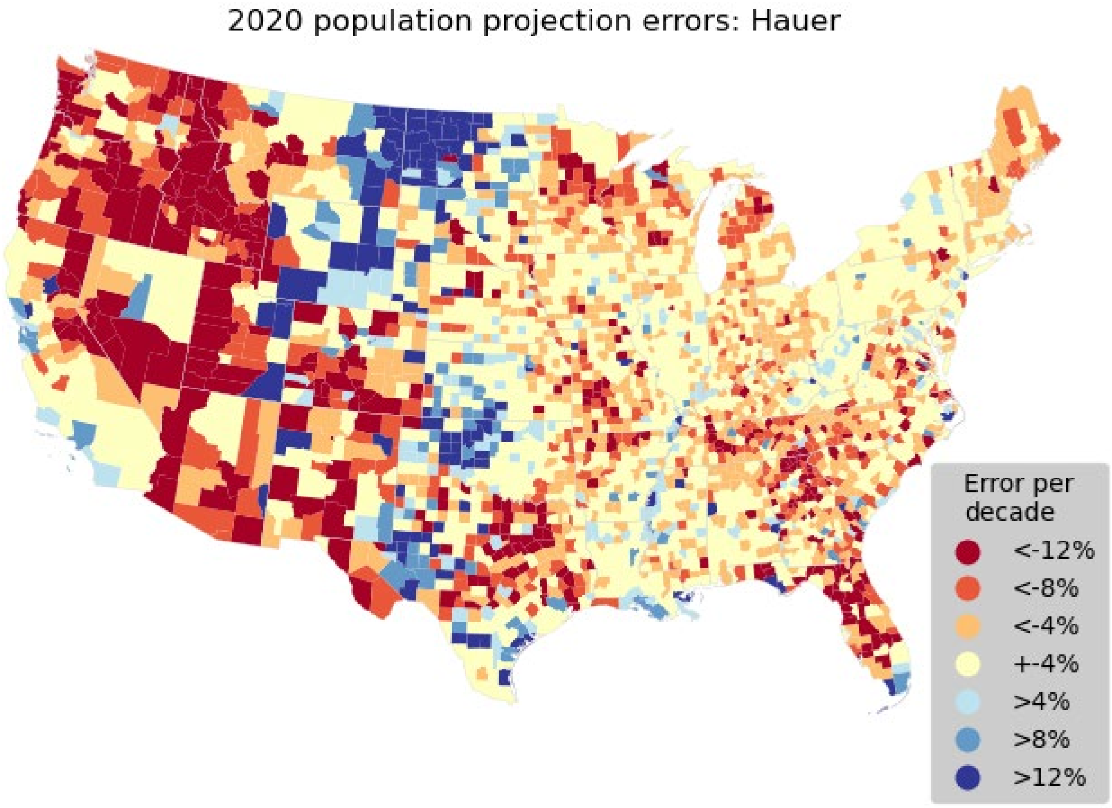

Figure 25. County level projection errors of Hauer (2019)

REFERENCES

Akwawua, S., & Pooler, J. A. (2000). An Intervening Opportunities Model of U.S. Interstate Migration Flows. *Geography Research Forum*, *20*, 33–51.

Allen, M. R., Fernandez, S. J., Fu, J. S., & Olama, M. M. (2016). Impacts of climate change on sub-regional electricity demand and distribution in the southern United States. *Nature Energy*, *1*(8), 16103. https://doi.org/10.1038/nenergy.2016.103

Azose, J. J., Ševčíková, H., & Raftery, A. E. (2016). Probabilistic population projections with migration uncertainty. *Proceedings of the National Academy of Sciences*, *113*(23), 6460–6465. https://doi.org/10.1073/pnas.1606119113

Beaujouan, É., & Sobotka, T. (2017). *Late motherhood in low-fertility countries: Reproductive intentions, trends and consequence* (Human Fertility Database Research Report HFD RR-2017-002). Vienna Institute of Demography. https://www.econstor.eu/handle/10419/156319

Bierwagen, B. G., Theobald, D. M., Pyke, C. R., Choate, A., Groth, P., Thomas, J. V., & Morefield, P. (2010). National housing and impervious surface scenarios for integrated climate impact assessments. *Proceedings of the National Academy of Sciences*, *107*(49), 20887–20892.

Bohara, A. K., & Krieg, R. G. (1996). A Zero-inflated Poisson Model of Migration Frequency. *International Regional Science Review*, *19*(3), 211–222. https://doi.org/10.1177/016001769601900302

Bowen, W. M., & Haynes, K. E. (2000). The Debate over Environmental Justice. *Social Science Quarterly*, *81*(3), 892–894.

Bowen, W. M., Salling, M. J., Haynes, K. E., & Cyran, E. J. (1995). Toward Environmental Justice: Spatial Equity in Ohio and Cleveland. *Annals of the Association of American Geographers*, *85*(4), 641–663. https://doi.org/10.1111/j.1467-8306.1995.tb01818.x

Burger, M., van Oort, F., & Linders, G.-J. (2009). On the Specification of the Gravity Model of Trade: Zeros, Excess Zeros and Zero-inflated Estimation. *Spatial Economic Analysis*, *4*(2), 167–190. https://doi.org/10.1080/17421770902834327

Cebula, R. J., Kohn, R. M., & Vedder, R. K. (1973). Some determinants of interstate migration of blacks, 1965-1970. *Economic Inquiry*, *11*(4), 500–505. https://doi.org/10.1111/j.1465-7295.1973.tb00979.x

Centers for Disease Control and Prevention. (2020a). *Underlying Cause of Death 1999-2019 on CDC WONDER Online Database*. http://wonder.cdc.gov/ucd-icd10.html

Centers for Disease Control and Prevention. (2020b, October). *Natality public-use data 2007-2019 on CDC WONDER Online Database*. https://wonder.cdc.gov/natality-current.html

Clark, W. A. V., & Onaka, J. (1983). Life Cycle and Housing Adjustment as Explanations of Residential Mobility. *Urban Studies*, *20*(1), 47–57. https://doi.org/10.1080/00420988320080041

Cromartie, J. (2020). *Rural America at a Glance: 2020 Edition* (Economic Information Bulletin No. 221; p. 6). United States Department of Agriculture, Economic Research Service.

DaVanzo, J. (1981). Repeat migration, information costs, and location-specific capital. *Population and Environment*, *4*(1), 45–73. https://doi.org/10.1007/BF01362575

Fik, T., Amey, R., & Mulligan, G. (1992). Labor migration amongst hierarchically competing and intervening origins and destinations. *Environment and Planning A*, *24*, 1271–1290.

Fik, T., & Mulligan, G. (1990). Spatial Flows and Competing Central Places: Towards a General Theory of Hierarchical Interaction. *Environment and Planning A*, *22*(4), 527–549. https://doi.org/10.1068/a220527

Fik, T., & Mulligan, G. (1998). Functional Form and Spatial Interaction Models. *Environment and Planning A*, *30*(8), 1497–1507. https://doi.org/10.1068/a301497

Flowerdew, R., & Aitkin, M. (1982). A method of fitting the gravity model based on the Poisson distribution. *Journal of Regional Science*, *22*(2), 191–202.

Flowerdew, R., & Lovett, A. (1988). Fitting Constrained Poisson Regression Models to Interurban Migration Flows. *Geographical Analysis*, *20*(4), 297–307. https://doi.org/10.1111/j.1538-4632.1988.tb00184.x

Fotheringham, A. S. (1983). A new set of spatial-interaction models: The theory of competing destinations. *Environment and Planning A*, *15*, 15–36.

Fotheringham, A. S., Brunsdon, C., & Charlton, M. (2000). Spatial Modelling and the Evolution of Spatial Theory. In *Quantitative Geography: Perspectives on Spatial Data Analysis* (p. 270). SAGE Publications.

Fotheringham, A. S., & Williams, P. A. (1983). Further Discussion on the Poisson Interaction Model. *Geographical Analysis*, *15*(4), 343–347. https://doi.org/10.1111/j.1538-4632.1983.tb00792.x

Fowler, C. S., & Jensen, L. (2020). Bridging the gap between geographic concept and the data we have: The case of labor markets in the USA. *Environment and Planning A: Economy and Space*, *52*(7), 1395–1414. https://doi.org/10.1177/0308518X20906154

Frey, W. H., & Liaw, K.-L. (2005). Migration within the United States: Role of Race-Ethnicity. In G. Burtless & J. Rothenburg Pack (Eds.), *Brookings-Wharton Papers on Urban Affairs 2005* (pp. 207–262). Brookings Institution Press. http://www.jstor.org/stable/25067420

Freymeyer, R. H., & Ritchey, P. N. (1985). Spatial Distribution of Opportunities and Magnitude of Migration: An Investigation of Stouffer's Theory. *Sociological Perspectives*, *28*(4), 419–440. https://doi.org/10.2307/1389227

Galle, O. R., & Taeuber, K. E. (1966). Metropolitan Migration and Intervening Opportunities. *American Sociological Review*, *31*(1), 5. https://doi.org/10.2307/2091275

Gao, J. (2017). *Downscaling Global Spatial Population Projections from 1/8-degree to 1-km Grid Cells* (NCAR Technical Notes NCAR/TN-537+STR). National Center for Atmospheric Research.

Gholdin, H. M. (1973). Kinship Networks in the Migration Process. *International Migration Review*, *7*(2), 163–175.

Gibson, J. G. (1975). THE INTERVENING OPPORTUNITIES MODEL OF MIGRATION: A CRITIQUE. *Socio-Economic Planning Sciences*, *9*(5), 205–208.

Granberry, P., Kim, C., Resseger, M., Lee, J., Lima, A., & Kang, K. (2018). Who Is At Risk of Migrating? Developing Synthetic Populations to Produce Efficient Domestic Migration Rates Using the American Community Survey. *Urban Science*, *2*(3), 80. https://doi.org/10.3390/urbansci2030080

Guldmann, J.-M. (1999). *Competing destinations and intervening opportunities interaction models of inter-city telecommunication flows*. *78*, 179–194.

Hardy, R. D., & Hauer, M. E. (2018). Social vulnerability projections improve sea-level rise risk assessments. *Applied Geography*, *91*, 10–20. https://doi.org/10.1016/j.apgeog.2017.12.019

Hauer, M. E. (2017). Migration induced by sea-level rise could reshape the US population landscape. *Nature Climate Change*, *7*(5), 321–325. https://doi.org/10.1038/nclimate3271

Hauer, M. E. (2019). Population projections for U.S. counties by age, sex, and race controlled to shared socioeconomic pathway. *Scientific Data*, *6*, 190005. https://doi.org/10.1038/sdata.2019.5

Hauer, M. E., Evans, J. M., & Alexander, C. R. (2015). Sea-level rise and sub-county population projections in coastal Georgia. *Population and Environment*, *37*(1), 44–62. https://doi.org/10.1007/s11111-015-0233-8

Hauer, M. E., Evans, J. M., & Mishra, D. R. (2016). Millions projected to be at risk from sea-level rise in the continental United States. *Nature Climate Change*, *6*(7), 691–695. https://doi.org/10.1038/nclimate2961

Haynes, K. E., Poston, D. L., & Schnirring, P. (1973). Intermetropolitan Migration in High and Low Opportunity Areas: Indirect Tests of the Distance and Intervening Opportunities Hypotheses. *Economic Geography*, *49*(1), 68. https://doi.org/10.2307/142746

Hildner, K. F., Nichols, A., & Martin, S. (2015). *Methodology and Assumptions for the Mapping America’s Futures Project*.

Hu, P., & Pooler, J. (2002). An empirical test of the competing destinations model. *Journal of Geographical Systems*, *4*(3), 301–323.

Internal Revenue Service. (2016). *2013-2014 IRS Migration Data Users Guide*.

Islam, S., Gandhi, D., Elarde, J., Anderson, T., Roess, A., Leslie, T. F., Kavak, H., & Züfle, A. (2021). Spatiotemporal prediction of foot traffic. *Proceedings of the 5th ACM SIGSPATIAL International Workshop on Location-Based Recommendations, Geosocial Networks and Geoadvertising*, 1–8. https://doi.org/10.1145/3486183.3490997

Jiang, L., O’Neill, B. C., Zoraghein, H., & Dahlke, S. (2020). Population scenarios for U.S. states consistent with shared socioeconomic pathways. *Environmental Research Letters*, *15*(9), 094097. https://doi.org/10.1088/1748-9326/aba5b1

Jones, B., & O’Neill, B. C. (2013). Historically grounded spatial population projections for the continental United States. *Environmental Research Letters*, *8*(4). https://doi.org/10.1088/1748-9326/8/4/044021

Kavak, H., Kim, J.-S., Crooks, A., Pfoser, D., Wenk, C., & Züfle, A. (2019). Location-Based Social Simulation. *Proceedings of the 16th International Symposium on Spatial and Temporal Databases*, 218–221. https://doi.org/10.1145/3340964.3340995
Lambert, D. (1992). Zero-Inflated Poisson Regression, with an Application to Defects in Manufacturing. *Technometrics*, *34*(1), 1. https://doi.org/10.2307/1269547

Lansing, J. B., & Mueller, E. (1967). *The Geographic Mobility of Labor*. Institute for Social Research.

Lee, E. (1966). A Theory of Migration. *Demography*, *3*(1), 47–57.

Lee, R., & Mason, A. (2014). Is low fertility really a problem? Population aging, dependency, and consumption. *Science*, *346*(6206), 229–234. https://doi.org/10.1126/science.1250542

Lo, L. (1992). Destination Interdependence and the Competing-Destinations Model. *Environment and Planning A*, *24*(8), 1191–1204. https://doi.org/10.1068/a241191

Martin, S., Nichols, A., & Hildner, K. F. (2017). *Methodology and Assumptions for the Mapping America's Futures Project*.

McGranahan, D. (1999). *Natural Amenities Drive Rural Population Change* (Agricultural Economic Report No. 781). Food and Rural Economics Division, Economic Research Service, U.S. Department of Agriculture.

McGranahan, D. A. (2008). Landscape influence on recent rural migration in the U.S. *Landscape and Urban Planning*, *85*(3–4), 228–240. https://doi.org/10.1016/j.landurbplan.2007.12.001

McGranahan, D., Cromartie, J., & Wojan, T. (2010). *Nonmetropolitan Outmigration Counties* (Economic Research Report No. 107).

McHugh, K. E. (1988). Determinants of Black Interstate Migration, 1965-70 and 1975-80. *Annals of Regional Science*, *22*(1), 36–48.

McKee, J. J., Rose, A. N., Bright, E. A., Huynh, T., & Bhaduri, B. L. (2015). Locally adaptive, spatially explicit projection of US population for 2030 and 2050. *Proceedings of the National Academy of Sciences*, *112*(5), 1344–1349. https://doi.org/10.1073/pnas.1405713112

McPherson, M., Smith-Lovin, L., & Cook, J. M. (2001). Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, *27*(1), 415–444.

Mele, A. (2021). A Structural Model of Homophily and Clustering in Social Networks. *Journal of Business & Economic Statistics*, 1–13. https://doi.org/10.1080/07350015.2021.1930013
Miller, E. (1972). A note on the role of distance in migration: Costs of mobility versus intervening opportunities. *Journal of Regional Science*, *12*(3), 475–478.

Murakami, D., & Yamagata, Y. (2019). Estimation of Gridded Population and GDP Scenarios with Spatially Explicit Statistical Downscaling. *Sustainability*, *11*(7), 2106. https://doi.org/10.3390/su11072106

Myrskylä, M., Kohler, H.-P., & Billari, F. C. (2009). Advances in development reverse fertility declines. *Nature*, *460*(7256), 741–743. https://doi.org/10.1038/nature08230

Newbold, K. B. (2014). *Population Geography: Tools and Issues* (Second Edition). Rowman & Littlefield.

O'Hara, R. B., & Kotze, D. J. (2010). Do not log-transform count data. *Methods in Ecology and Evolution*, *1*(2), 118–122. https://doi.org/10.1111/j.2041-210X.2010.00021.x

Ortman, J. M., Velkoff, V. A., & Hogan, H. (2014). *An Aging Nation: The Older Population in the United States* (No. P25-1140; Current Population Reports). U.S. Census Bureau.

Pack, J. R. (1973). Determinants of Migration to Central Cities. *Journal of Regional Science*, *13*(2), 249–260.

Pellegrini, P. A., & Fotheringham, A. S. (1999). Intermetropolitan Migration and Hierarchical Destination Choice: A Disaggregate Analysis from the US Public Use Microdata Samples. *Environment and Planning A*, *31*(6), 1093–1118. https://doi.org/10.1068/a311093

Piovani, D., Arcaute, E., Uchoa, G., Wilson, A., & Batty, M. (2018). Measuring accessibility using gravity and radiation models. *Royal Society Open Science*, *5*(9), 171668. https://doi.org/10.1098/rsos.171668

Plane, D. A., & Heins, F. (2003). Age articulation of US inter-metropolitan migration flows. *The Annals of Regional Science*, *37*(1), 107–130.

Plane, D. A., Henrie, C. J., & Perry, M. J. (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.

Plane, D. A., & Jurjevich, J. R. (2009). Ties that no longer bind? The patterns and repercussions of age-articulated migration. *The Professional Geographer*, *61*(1), 4–20.

Price, D. O. (1971). Rural to Urban Migration of Mexican Americans, Negroes and Anglos. *International Migration Review*, *5*(3), 281–291.

Raphael, S. (1998). Intervening Opportunities, Competing Searchers, and the Intrametropolitan Flow of Male Youth Labor. *Journal of Regional Science*, *38*(1), 43–59. https://doi.org/10.1111/0022-4146.00081

Ravenstein, E. G. (1889). The Laws of Migration. *Journal of the Royal Statistical Society*, *52*(2), 241. https://doi.org/10.2307/2979333

Rogerson, P. A. (1986). Parameter estimation in the intervening opportunities model. *Geographical Analysis*, *18*(4), 357–360.

Sanderson, W. C. (1998). Knowledge Can Improve Forecasts: A Review of Selected Socioeconomic Population Projection Models. *Population and Development Review*, *24*, 88–117. https://doi.org/10.2307/2808052

Sanstad, A. H., Johnson, H., Goldstein, N., & Franco, G. (2009). *Long-run socioeconomic and demographic scenarios for California* (CEC-500-2009-013-F; The California Climate Change Center Report Series, p. 49). California Climate Change Center.

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

Silva, J. S., & Tenreyro, S. (2010). On the existence of the maximum likelihood estimates in Poisson regression. *Economics Letters*, *107*(2), 310–312. https://doi.org/10.1016/j.econlet.2010.02.020

Silva, J. S., & Tenreyro, S. (2011). Further simulation evidence on the performance of the Poisson pseudo-maximum likelihood estimator. *Economics Letters*, *112*(2), 220–222. https://doi.org/10.1016/j.econlet.2011.05.008

Simini, F., González, M. C., Maritan, A., & Barabási, A.-L. (2012). A universal model for mobility and migration patterns. *Nature*, *484*(7392), 96–100. https://doi.org/10.1038/nature10856

Smith, S. K., Tayman, J., & Swanson, D. A. (2013). *A Practitioner's Guide to State and Local Population Projections*. Springer.

Spring, A., Tolnay, S. E., & Crowder, K. (2016). Moving for Opportunities? Changing Patterns of Migration in North America. In W. J. Michael (Ed.), *International Handbook of Migration and Population Distribution* (Vol. 6, pp. 421–448). Springer.
Stewart, J. Q. (1941). An Inverse Distance Variation for Certain Social Influences. *Science*, *93*(2404), 89–90.

Stouffer, S. A. (1940). Intervening Opportunities: A Theory Relating Mobility and Distance. *American Sociological Review*, *5*(6), 845. https://doi.org/10.2307/2084520

Stouffer, S. A. (1960). Intervening Opportunities and Competing Migrants. *Journal of Regional Science*, *2*(1), 1–26. https://doi.org/10.1111/j.1467-9787.1960.tb00832.x

Tofallis, C. (2015). A better measure of relative prediction accuracy for model selection and model estimation. *Journal of the Operational Research Society*, *66*(8), 1352–1362. https://doi.org/10.1057/jors.2014.103

Tornqvist, L., & Vartia, P. (1985). How Should Relative Changes Be Measured? *The American Statistician*, *39*(1), 43–46.

University of Virginia. (2016). *State and National Projections Methodology*. Weldon Cooper Center for Public Service.

U.S. Census Bureau. (1990). *Enhanced Migration Files, v1 (1985–1990)*. Socioeconomic Data and Applications Center. http://dx.doi.org/10.7927/H4057CV3

U.S. Census Bureau. (2000). *County-to-County Migration Flow Files*. Census 2000 Gateway. https://www.census.gov/population/www/cen2000/ctytoctyflow/index.html

U.S. Census Bureau. (2014a). *2009-2013 ACS Migration Flow Files Documentation*. http://www.census.gov/content/dam/Census/topics/population/migration/guidance-for-data-users/acs-migration-tutorial/2009-2013-Migration-Flows-Documentation.pdf

U.S. Census Bureau. (2014b). *Methodology, Assumptions, and Inputs for the 2014 National Projections*.

U.S. Census Bureau. (2021). *Annual County Resident Population Estimates by Age, Sex, Race, and Hispanic Origin: April 1, 2010 to July 1, 2020*. https://www2.census.gov/programs-surveys/popest/technical-documentation/file-layouts/2010-2020/cc-est2020-alldata.pdf

U.S. Energy Information Administration. (2017). *Annual Energy Outlook 2017 with projections to 2050* (No. AEO2017).

U.S. Environmental Protection Agency. (2009). *Land-Use Scenarios: National-Scale Housing-Density Scenarios Consistent with Climate Change Storylines* (EPA/600/R-08/076F). U.S. Environmental Protection Agency (EPA).
U.S. Environmental Protection Agency. (2016). *Updates to the Demographic and Spatial Allocation Models to Produce Integrated Climate and Land Use Scenarios (ICLUS) Version 2* (EPA/600/R-16/366F). National Center for Environmental Assessment.

U.S. Global Change Research Program. (2015). *Towards Scenarios of U.S. Demographic Change: Workshop Report* [Workshop Report].

Viboud, C., Bjornstad, O. N., Smith, D. L., Simonsen, L., Miller, M. A., & Grenfell, B. T. (2006). Synchrony, Waves, and Spatial Hierarchies in the Spread of Influenza. *Science*, *312*(5772), 447–451. https://doi.org/10.1126/science.1125237

Wadycki, W. J. (1975). Stouffer's Model of Migration: A Comparison of Interstate and Metropolitan Flows. *Demography*, *12*(1), 121. https://doi.org/10.2307/2060737

Wear, D. N., & Prestemon, J. P. (2019). Spatiotemporal downscaling of global population and income scenarios for the United States. *PLOS ONE*, *14*(7), 1–19. https://doi.org/10.1371/journal.pone.0219242

Weeks, J. R. (2012). *Population: An Introduction to Concepts and Issues* (11th ed.). Wadsworth.

Wittgenstein Centre for Demography and Global Human Capital. (2018a). *Age-Specific Fertility Rate*. https://www.wittgensteincentre.org/dataexplorer

Wittgenstein Centre for Demography and Global Human Capital. (2018b). *Age-Specific Survival Ratio*. https://www.wittgensteincentre.org/dataexplorer

Xu, Z. (2017). The structure and dynamics of population migration among economic areas in the United States from 1990 to 2011: US inter-area population migration networks. *Papers in Regional Science*. https://doi.org/10.1111/pirs.12282

Zarnoch, S. J., Cordell, H. K., Betz, C. J., & Langner, L. (2010). *Projecting county-level populations under three future scenarios: A technical document supporting the Forest Service 2010 RPA Assessment* (General Technical Report SRS-128; p. 41). U.S. Department of Agriculture, Forest Service, Southern Research Station. http://www.treesearch.fs.fed.us/pubs/35892

Zelinksy, W. (1971). The Hypothesis of the Mobility Transition. *The Geographical Review*, *61*, 219–249.

Zipf, G. K. (1946). The P_1P_2/D Hypothesis: On the Intercity Movement of Persons. *American Sociological Review*, *11*(6), 677. https://doi.org/10.2307/2087063

BIOGRAPHY

Philip Morefield graduated from Westside High School, Omaha, Nebraska, in 1996. He received his Bachelor of Science from the University of Nebraska at Omaha in 2004, and his Master of Environmental Management from Portland State University in 2008. He has been a researcher at the U.S. Environmental Protection Agency since 2010.