

Why Can’t They Just Look It Up? Utilizing Restricted Administrative Data to Overcome the Limitations of Surveys in Demography

Paul Scholes

3rd Year PhD Student

Texas A&M Sociology

Projected Dissertation Proposal Defense: Fall 2025

Disaster-related migration is hard to measure. Research “often rel(ies) either on census or survey data” (Berlemann and Steinhardt 2017). Censuses only occur rarely, and the intervals are often too large to differentiate between migration from disasters from migration for other reasons. Administrative records capture demographic shifts due to disasters, including deaths, migration, and staying. This project uses administrative data to further our knowledge of disaster-related migration. Beyond migration, this approach can improve demographic estimates, like life expectancy or fertility, particularly for populations underrepresented in traditional surveys

Surveys and other probability-based data sources are often used to generate inferences for a population. However, there are drawbacks to this approach and some of them are growing more consequential. As described by leading economists, “the research frontier moves to use administrative data” (Card et al. 2010:1) for a couple of reasons. First is the cost of sampling and gathering data, which is already paid for in administrative data. Second (and one of the reasons that primary data collection has gotten so expensive recently) are the continually declining response rates in recent years. Non-respondents may be systematically different than those who respond, for example in surveys utilizing phone number sampling frames, there is often a dearth of young and/or poor people, which bias population estimates (Ambel, McGee, and Tsegay 2021).

This study demonstrates how administrative data can enhance demographic analysis by creating a comprehensive migration frame that does not rely on traditional surveys. I have already contributed as a part of a team to (1) create a dataset using unique identifiers for people maintained by the U.S. Census Bureau called Primary Identification Keys (PIKs) and identifiers for addresses called Master Address Filer Identifiers (MAFIDs). (2) I will argue for the frame’s validity by illustrating PIKs coverage, using the American Community Survey and the 2020 Decennial Census as separate comparators. (3) I will observe the comparability of this new dataset — a demographic frame— made of administrative records, to the ACS and decennial census in 2020 by examining the coverage error for each geography in these datasets. Having shown the ability to make estimates with the demographic frame, I will create a similar, but entirely self-directed approach to identify movers and then (4) estimate measures of migration, such as an in/out migration matrix and overall migration efficiency, with these three data sources and discuss coverage differences across different geographies. While data is available at various geographic granularities, disclosure review will determine the geographic level statistics are presented in. Finally, I will (5) use these datasets to compare the migration related to hurricane Ian.

Disaster migration theories have primarily analyzed individual-level decisions based on push/pull factors (Lee 1966) mitigating risk (Stark and Taylor 1991) and responses to social networks (Massey 2015). Migration matrices allows theory to abstract to new aggregations, such as geography or housemates, and observe processes unseen by other methods. For example, methods that cannot differentiate between deaths, survey attrition, and migration.

Getting access to data will be the hardest part for other researchers. I have access because I have been working on the Census Bureau prototype administrative data frame – the Demographic Frame— for several years now. While the Demographic Frame is an important data source that utilizes administrative data to circumvent the drawbacks of surveys, its value is limited to timeframes with an extract. A Business Rules Approach to Person/Place Matching (BRAPPA) is built from the knowledge generated from this team, but (1) is being assembled from the ground up. It will be used to update a extract or other data source to any reference date by referencing a corpus of administrative records. The demographic frame utilizes a modeling strategy to match PIKs with MAFIDs, while (2) this system uses no modeling at all, only programmatic logic. The goal of the demographic frame is to provide a frame for the whole country that researchers can use easily and flexibly, combining records across several years for a any reference date. (3) This project creates a person/place data frame that researchers can assemble and modify the logic to tailor the assumptions they make. This is more labor intensive but allows researchers greater freedom to design studies. Freedom to choose their own reference dates and sources to identify movers and non-movers.

Current work on disasters often utilize a single unrepresentative data source, like twitter users (Zou et al. 2019), or hospital records (Craig et al. 2013, 2018). Current work on migration often use a single administrative dataset like the Internal Revenue Service records or the American Community Survey (Molloy, Smith, and Wozniak 2011). This project will supplement existing methods by describing the context of a time or place by leveraging available administrative data from hundreds of sources. These administrative datasets need to be combined to make a data frame to answer these needs. I will make this data frame, compare it to other methods, and then use hurricane Ian to apply it to disaster migration. This study advances demographic methodology and disaster migration theory by introducing a replicable framework for estimating migration trends. By enabling the systematic study of small and vulnerable populations, this approach enhances both theoretical insights and policy responses to disasters. It also contributes to the efforts of the U.S. Census Bureau’s Demographic Frame and illustrates new applications to its innovative approach.

Chapter 1: Data and Strategies for Person Place Assignment

BACKGROUND

*“Migration scholars of today generally have to make two decisions to define migrants: 1) they choose geographic units to define potential origin and destination locations; and 2) they define the time period in which individuals move between origin and destinations.” (Molloy, Smith, and Wozniak 2011:175)*

This project introduces several new or emerging ways to analyze the migration of people to locations across the United States. I start by discussing the central elements of migration: people, places, and time. I then turn to a review of current strategies and their limitations before discussing the new proposed approaches.

*Operationalization of Geography*

Researchers have tended to choose potential origin and destination locations based on the availability of data and the theoretical orientations held. Many migration researchers leverage U.S. Census Bureau data for addresses for a few reasons. The Census Bureau is the only organization that attempts to enumerate every person in the United States at the place they live, and the only organization hosting and maintaining the MAF-Tiger system, which combines geospatial and housing information together in a single system. Another popular option is to analyze surveys, but even surveys like the Current Population Survey, the American Community Survey, or the Survey of Income and Program Participation use Census Bureau data to make sampling frames or to manage and harmonize the addresses of respondents (United States Census Bureau 2019, 2025g, 2025a).

For U.S. based migration, many researchers use economic regions often called metropolitan statistical areas or core-based statistical areas. These areas are built using counties or county equivalents by the U.S. Office of Management and Budget (United States Census Bureau 2025b). The U.S. Census Bureau uses much smaller geographies, like tracts, blocks and block groups, but these are designed to reflect county, state geographies whenever possible. The ability for smaller geographies to nest into these larger geographies allows for smoother aggregation to larger levels of measurement.

Problems arise when trying to aggregate (or disaggregate) into boundaries with messy boundary overlays. For example, because zip codes often straddle county and state lines, it can be hard to know how the characteristics of a county are reflected in a zip code which contains multiple counties. This incongruence between aggregations is called the modifiable areal unit problem and, when ignored, leads to the ecological fallacy or the assumption that characteristics of an aggregate will hold for individuals and vice versa.

The nesting of Census Bureau geographies ensures that estimates have the same geographical basis as data is aggregated or disaggregated. However, clean nesting of geographies does not address the arbitrary, or sometimes instrumental, desires of the people and organizations who drew the boundaries. This portion of the modifiable areal problem, the goals and whims of the boundary makers, “is ever-present” (Buzzelli 2020:1).

Migration researchers often use individual level data where a respondent’s geography is noted and used in modeling. This approach is widely used, especially in multilevel modeling (Garson 2019; Khaw et al. 2021; Zhou et al. 2022). There are several drawbacks to this approach: 1) Data anonymity is difficult to preserve unless the researcher is only presenting estimates, additionally giving researchers access to respondent’s addresses may be a problem: either for respondents or ethical guidelines. 2) There may not be enough data on each geography to obtain reasonable estimates 3) The processing time for modeling individual level characteristics accounting for the multi-level nature of these interactions often requires a lot of computing power, even to the point of being impossible to estimate (Garson 2019).

Federal data centers address many of these geographic, ethical, and computational problems. Researchers can access microdata, which is not aggregated at all and thus theoretically sidesteps the modifiable areal problem. However, these researchers are usually limited by disclosure review boards. Because of the anonymity concerns in identifying individual migrants across places, analyses are aggregated up to a larger geography, and because of the modifiable areal problem and the nesting of smaller Census Bureau geographies into counties and state, these aggregations are usually counties or states. This protects individual respondents and side steps the imperfect nesting of geographies portion of the modifiable areal unit problem. Thus, a popular strategy is to use individual level data when possible and aggregate up to larger geographies when needed.

Identifying an aggregation strategy is not the only geographic consideration, the places people live, usually housing units located at addresses, require careful thought as well. Migration projects need to enumerate where people live; and people can live in buildings or non-conventional housing like boats, railroad cars, tents, or vehicles, so identifying the suis generis unit of analysis is not simple.

The Census Bureau maintains a database called the Master Address File (MAF), which identifies unique structures and single units within a structure (called MAF units) by identifiers called MAFIDs. MAF Unites represent residential, non-residential structures, and their addresses, and thus a micro-level geographic source. Many projects examining migration in the U.S. use MAF units and many data sources, like the American Community Survey, the Survey of Income and Program Participation, or the Current Population Survey, are using MAF units behind the scenes for their sampling frames and address reconciliation system (United States Census Bureau 2019, 2025g, 2025a). The MAF is updated frequently from state, federal (including the United States Postal Service), and commercial sources (usually for land parcel data). While the Geography Division continuously maintains the MAF, most users (internal and external to the Census Bureau) use a semi-annual extract called the MAFX.

The MAFX helpfully provides characteristics for MAF units, but MAF units are not necessarily places where people live. For this, the Census Bureau defines housing units. A housing unit is “a house, apartment, a mobile home or trailer, a group of rooms, or a single room (that can be temporarily vacant but intended for occupancy as a sperate living quarters).”(United States Census Bureau 2021:1).

There are two additional criteria: separateness, or living separate from others in the building, and direct access, the ability to enter the housing unit from the outside or a common hall. Finally non-conventional housing is included in the housing unit inventory when they are the usual place of residence for a person, otherwise they are excluded. Different surveys and programs use different thresholds for considering a housing unit, and researchers who are interacting with individual-level data.

Housing units are assigned MAFIDs, and these are used to anonymize data so researchers can analyze the people connected to the MAF units and the MAF units themselves without putting residents at risk. The difficulty for data users is that while all housing units are MAF units, MAF units are not necessarily valid housing units. How many (if any) individuals or families live in MAF unit would be unclear if the Census Bureau did not provide a mechanism to attach people to their residences. I will discuss more about the matching process in a section on administrative data.

*Operationalization of People*

Places are only half of the migration analysis, I now turn to people. Like the concerns around privacy for MAF units and addresses, the Census Bureau provides a system to anonymize the people that live in the United States: The Primary Identification Key or PIK. The identifiers are non-informative (no information can be gleaned from the number itself), anonymous, and durable over time. They are assigned probabilistically using the Census Bureau’s access to many administrative records, like the Numident and IRS tax payer data (Alexander and Genadek 2023). Everyone with a social security number or an individual taxpayer identification number is assigned a PIK and then the Census Bureau’s Person Identification Validation System matches the characteristics of people recorded in administrative records (such as names, dates of birth, birthplaces, and/or residences) with their most probable PIKs.

We now turn our focus from geography to demography, the people who live in the places. Like the concerns around privacy for MAF units and addresses, the Census Bureau provides a system to anonymize the people that live in the United States: The Primary Identification Key or PIK. The identifiers are non-informative (no information can be gleaned from the number itself), anonymous, and durable over time.

PIKs were first created to match people across older decennial censuses, but were quickly adopted for identifying and anonymizing living people today. The Census Bureau assigns PIKs based on a probabilistic matching model that incorporates available information—like first and last names, sex, residential addresses, and ages— from the administrative record when available. Information is not always available on every administrative record, many, like birth certificates, for example, do not include addresses which complicates the process and a probability is assigned across the various fields. NORC validated the PIK assignment process and found that the Census Bureau correctly matches characteristics to the proper person about 90% of the time across various representative samples (Mulrow et al. 2023). There are very small spatial patterns, the Southwest states have slightly lower match rates than the CEF, off by no more than 3% and usually around 1%. There is also a socioeconomic component. African Americans, poorer people, less educated, unemployed, uninsured, and those who spoke a language other than English at home were slightly less likely to be matched correctly. Despite these limitations, indeed many surveys and sources have similar limitations, Census Bureau products with PIKs are frequently used in sociological analysis (Alvaredo et al. 2013; Liebler et al. 2017; Piketty and Saez 2014; Porter 2008).

People changing their characteristics, like names, can make it difficult to match records collected for different purposes. U.S. Census Bureau has had a whole division working on the matching problem for many years (Brittany Bond et al. 2014) and for modern records, largely overcoming it. PIKs cover about 2.5% fewer people than reported in the 2020 Decennial Census and about 1.8% fewer people than in the official 2020 population estimates (Ortman and Knapp 2023). The false match rate for the Person Identification Validation System is around .005% (Layne, Wagner, and Rothhaas 2014).

*Operationalization of Time*

Researchers also consider the time frames used for migration research. Migration takes a continuous phenomenon and, out of computational necessity, makes discrete periods of time to analyze the migration in. Typically, the discrete cuts of time are a year, or several years. But this creates measurement issues when migration is temporary or rapid. People may move multiple times between data collection periods or move to a new location and then back again.

The gold standard would be knowing the day a person moves, but most data, especially administrative data, only registers a change when a new vintage of data is collected. The benefit of access to many administrative data sources is that the observation opportunities increase with the number of sources. Individuals who file taxes have observation windows around April, while Selective Service records are updated when a male turns 18 (and registers for the draft).

Many records do not have dates attached to them. When this happens, the applicable date for the record is imputed according to characteristics known about the source. For example, many state Supplemental Nutrition Assistance Program (SNAP) records have the month the participant files for benefits, but not the day or the year. The year can be imputed from the vintage of the dataset, i.e. the 2016 delivery contains 2016 data. In this case, imputation for the day follows similar logic as imputation in other demographic variables. For demographic imputations it is common to assume uncertain demographic events happen at the middle of the year (Wachter 2014) to avoid a systematic under or overcount. In a demographic context, researchers assume uncertain births and deaths happen in the middle of a year (with only a couple of exceptions) to. In a migration context, migration events are usually imputed at the middle of the missing value: e.g. the 15th of the month for a missing day, June for a missing month, etc.

*Data Sources Used for Migration Research*

*Surveys.* The defining characteristic of surveys is the selection of a subset of respondents from a sampling frame. Ideally the sampling frame is an enumeration of all the possible units of analysis, though this is rarely attainable. The American Community Survey and the Decennial Census operate using the MAFX as their sampling frame. (The Decennial Census visits every housing unit in the sampling frame and adds housing units as it discovers new ones). This address-focused approach is a major advantage over other sampling frames, like phone numbers. However, there are some drawbacks to an address-based sampling frame. Addresses are not people, but for many studies, people are the unit of analysis. Any relationship between sampling frame and unit of analysis can bias estimates. For example, the tendency for young and poor people to not have landlines has been a bias in several studies across the world (Ambel et al. 2021). Similarly, though to a lesser extent, sampling frames based on addresses will overrepresent richer people with many houses, and underrepresent those without addresses, those who move out of the country, or those who only fill out addresses on public forms with a P.O. box.

The alignment between sampling frame and dependent variable is also important to consider. Phone surveys that do not take into account person information tend to bias variables and outcomes related to phone owning, like age, health, or income (Ambel et al. 2021; Call et al. 2011; Gourlay et al. 2021). In migration research, both the people selected and the types of people who can move: wealthier, educated, and younger (Feliciano and Lanuza 2017; Stark and Taylor 1991), ideally need to be aligned to prevent bias.

The American Community Survey (ACS) is the successor of the Decennial long form. The long form had many questions, but pertinently asked “Where did you live five years ago?” (United States Census Bureau 2025f), which yielded five-year migration estimates at every decennial census. Without the long form, 10-year migration estimates are possible using the short form alone. Prior to 2010, 1 in 6 American households filled out the decennial long form. The long form is replaced by the ACS in 2010 and in 2011 3.57 million addresses (households) are sampled each year to create the ACS estimates (United States Census Bureau 2025b, see chapter 4). Like the long form, the ACS has many questions but pertinently asks “Did this person live in this house or apartment 1 year ago?” and “Where did this person live 1 year ago?” yielding one-year migration estimates for large areas with many households sampled. For small areas, ACS responses are aggregated into one-year estimates from a period of five years. For example, while one year migration rates for small counties can be estimated by combing five years of responses together, the question and subsequent estimates are still for one year.

There are some serious drawbacks to using ACS data for migration. The ACS data must be combined to get good sample sizes for many smaller counties and so county migration estimates are only available for non-overlapping five-year spans, e.g. 2011-2015, 2016-2020, etc.). Additionally, the ACS, like all surveys, has coverage errors, which represent the probability that a unit makes it into the sample more often than they should (over-coverage), or is less likely to be represented in the sample (under-coverage). The ACS suffers from over coverage and under coverage. In 2023, the ACS had a coverage rate of 83% for Black non-Hispanics and a coverage rate of 153.6% for Non-Hispanic American Indian and Alaskan Natives (United States Census Bureau 2025e).

The Decennial Census is a valiant effort of enumeration of all residents in the United States. In this sense, the Decennial Census covers the same population or universe as the ACS, but instead of surveying a sample of residents, enumerates all it can contact. Decennial enumeration is required by the U.S. constitution, and so sample based methods are legally prohibited. Residents sometimes fill out the Census dishonestly, incorrectly, or fail to comply. These are filtered out or imputed and published in the Census Edited File (CEF). While the Census Unedited File exists, is available, and is used for the official population counts, the CEF edits and imputes person characteristics like addresses, race, age, and sex (Devine, Jonathan, and Ryan 2021).

Despite the thousands of man-hours invested by enumerators, respondents, software, and internal analysts, the Decennial Census has major drawbacks in measuring migration. The largest drawback being its decennial nature, migration is not available for events, and many people will have moved more than once in a decade, which prevents researchers from getting an accurate picture of migration rates. Other concerns include struggles to accurately count the homeless, those off the grid, or sensitive populations like Native Americans on the reservations. Similar to the aims of this project, the 2020 Decennial Census used administrative records: IRS, Medicare/Medicaid, Household Composition File, and the Indian Health Service Patient Database to fill in the gaps (Mulry and Tello-Trillo 2023).

While surveys have been the dominant method for analyzing migration for nearly 200 years, there are several issues with surveys that have been growing more relevant and impactful with no signs of reversing: declining participation rates, increased respondent burden, and raising costs of surveys (Groves 2011).

Surveys that are relevant for researchers, including federal surveys, are often quite long and often ask sensitive or personal questions. While a common rule of thumb is to try to have the survey take no more than 10 minutes to complete, Federal surveys like the ACS can take much longer. Additionally, surveys often ask personal questions, and many of them, like the National Survey of Adolescent Health or the National Health and Nutrition Examination Survey take biological samples form respondent’s bodies. This affects participation rates and individual item response rates (Boyle et al. 2021). Developments like these mean that surveyors need to spend more time and money trying to balance their surveys and track down reluctant respondents.

Participation rates on surveys, both voluntary and mandatory, have been decreasing since the 1990’s (John Czajka and Amy Beyler 2016). The Current Population Survey is a voluntary survey that had response rates floating around 90% in 2010-2013, and response rates around 68% in 2025 (Bureau of Labor Statistics 2025). Even mandatory surveys, where participation is compelled by law, like the American Community Survey (ACS) see a decline in response rates. The ACS had a response rate of about 95% in the early 2000’s, while that rate has dropped to 82.9% in 2024 (United States Census Bureau 2025c). Every household that selects not to participate in a survey introduces selection effects into the design. These selection effects are expensive to address and introduce patterns into the data that a troublesome to resolve. Indeed, it is impossible to know how a sample differs from the general population without analyzing the whole population or making another sample that hopefully has better response rates/patterns. U.S. Census Bureau uses many smaller surveys to address coverage issues with the decennial census (US Census Bureau 2024).

*Administrative records.*

Administrative records are records like birth registries, tax filing information, or program enrollments. Administrative records are often used as the sampling frames for surveys, like the ACS and Decennial Census using the MAFX. These have valuable data on people in them, and many are hopeful that they can ease or eliminate modern surveys (Chun et al. 2021). The first and biggest issue with using administrative records is matching respondents across different records (Harron et al. 2017).

The Internal Revenue Service (IRS) is an administrative record frequently used in migration research (the other ubiquitous sources are the American Community Survey and the Decennial Census) (Hauer and Byars 2019). About 86% of the United States is represented in the county-to-county estimates published by the IRS (Molloy et al. 2011), about 116 million households. While the IRS data is released more frequently and has a much larger sample than the ACS, the IRS only examines households with income and lacks characteristics of the individual/household such as race, ethnicity, educational attainment, and more. Additionally, households may lie on their taxes to be taxed at a lower rate or other instrumental reasons.

With access to IRS, ACS, and Census data, many of the drawbacks of a particular data source can be ameliorated. Using PIKs and MAFIDs, characteristics that appear in one dataset like person-level characteristics in the Decennial Census or ACS, can be merged into more frequent or larger sample datasets like the IRS information, which lacks these characteristics. Additionally, we can leverage the information available in many more administrative records to increase coverage, for example including Bureau of Prison data to include the incarcerated population or Medicare data for the elderly. Administrative data can also resolve measurement errors from a single data source, because one-off mistakes will be ignored in favor of consensus of multiple sources.

*Third-party records.*

Third-party records are created by data brokers, like Acxiom or Eyeota, usually with the goal of profiling customers. This information is most often gathered through internet cookies, which while often required to make websites run correctly, also gather information about the interaction between website and user. Modern websites are multi layered, multi-creator spaces with third-party infrastructure doing advertising, social networking, analysis, and more (Mayer and Mitchell 2012). While third-party services are often collecting and selling user data, they are also responsible for many of the free content users enjoy. These exchanges between user and website partners have led to many innovations. There is broad but conditional public support for sharing personal information for third party or secondary use (Baines et al. 2024). Individuals value transparency, and control over the particulars of the exchange, including duration of access, what is shared, and who it is given to.

Anonymized, but matchable and trackable, user records are usually sold to other organizations and websites who do not have records on that user. Records can be aggregated and modeled into profiles to target various demographics such as age and sex, and various interests such as sports, or politics. The performance of these models is almost always unknown. It is rare for a data broker to reveal their strategies or allow independent parties to validate their data.

Another drawback to third party data is the increasing risk that individuals are doxed or reidentified by others. Researchers have broken the privacy safeguard of companies several times (Narayanan and Shmatikov 2007; Xin et al. 2025), and data leaks can make this problem worse. Companies may violate their own privacy policies when user information can be reidentified by other users, and while the U.S. and E.U. have punished some companies for ignoring their own privacy policies, enforcement and policy transparency has been modest but improving (Linden et al. 2019).

Despite these drawbacks, researchers often use third party data (Bronson, Lento, and Wiener 2015; Lathan et al. 2023; Markovikj et al. 2013), even in projects that require multiple time periods of data, such as disaster/forced migration (Böhme, Gröger, and Stöhr 2020; Zou et al. 2018).

The Pew Research Center was given one of the rare chances to validate 3rd party data in an analysis of voter records (Mitchell 2018). They link survey a sample of voters and compare their responses with five datasets from data brokers, whose identities are hidden. They find that data brokers are using a mix of administrative records, modeling, and third-party data like credit reports, or hunting magazine subscriptions to build their frames.

While data brokers often have access to correct information from these sources, the models often get these characteristics wrong anyway. For example, even when the correct race is listed in primary sources (like voting records) for 96% of the sample, the accuracy for the five datasets range from 74% to 85%. These accuracy rates are even worse for minorities, like accuracy ranging from 56% to 76% for African Americans, and 64%-75% for Hispanics. This example on race reveals a trend seen for religious affiliation or educational attainment: variables that clump around a single variable and/or have fewer levels tend to have higher accuracy than variables with more uniform spread and more levels.

If a person place assignment process could (1) correctly include PIKs for 80% of the population and (2) accurately predict PIK characteristics of that 80% sample about 80% of the time, it would be better than the best performing categories of the best performing third party data sets. This is the current frontier of whole-universe estimation, though no broker or administrative record can currently do both at the 80% level.

*The demographic frame extract.* The U.S. Census Bureau’s Demographic Frame (demoframe) Team’s extract uses administrative records and modeling to match people and places in a given time (Demographic Frame Team 2025). The demoframe extracts are modeled using people living in the United States known to the Social Security Administration (Anthony Wray, Board, and Administration 2024). The demoframe team uses the Census Bureau’s version of the Numident: the Census Numident.

The demoframe extract offers four machine learning models which identify the best PIK/MAFID pairs for a given year and reference date: an elastic net, random forest, logit, and boosted tree model. They are trained on the extract year’s ACS data as a truth set and then uses the sources from the past two years in the Person Place Table to create PIK/MAFID pairs. Other features used for training include the source name and the characteristics of a particular source, and the date a source was considered valid. Each model seems to have different strengths and weaknesses and there are different versions of the demoframe extracts with various reference dates and coverages. The assignment process and logic behind the PIK/MAFID pairs is opaque for each model. However, broadly speaking, the models prefer PIK/MAFID pairs with many corroborating sources, with higher quality sources, and more recent sources. It also has a feature which prefers mafids that were considered valid housing units during the last decennial census.

Internal analyses suggest using the random forest and logistic regression models over the elastic net and boosted tree models (Demographic Frame Team 2025). They suggest using either the random forest or logistic models based on their internal analyses, and so because the projected probability for the PIK/MAFID pairs are unimportant for this project, I use the random forest model whenever the demoframe extracts are used.

Having discussed administrative records, PIKs and PIK variable assignments, the next consideration for person place matching are the assignment of MAFIDs to MAF units in administrative records. Many administrative data sources, like the United States Postal Service, also have their own MAFID matching processes, which can introduce error. When MAFID assignment on administrative records use different versions of the MAFX, errors can accumulate, especially when the MAFX version year and the administrative record vintage are very different. The Census Bureau is the largest statistical agency of the United States, and thus links most MAFIDs in administrative datasets used by the Bureau. However, the Economic Research Division uses the most recent MAFX when assigning MAFIDs to the administrative records they process, while the records themselves are often not from the year of the MAFX used. For data that is delivered occasionally, the incongruences between the geography represented in the MAFX and the geography represented by the records can be substantial (Blaylock and Scholes 2025).

To address the issue of administrative records using different mafids that represent the same MAF unit, I created a table called the MAF\_Master. It combs through recent MAFXs to track MAFIDs, their characteristics, and their successors over time. The creation of this table is one of the novel contributions of the project (Blaylock and Scholes 2025).

The MAF\_Master contains information for all the MAFIDs seen in a MAFX from 2010 to the present. It contains information about which MAFIDs are purged from the MAF and when these purges occur. It also contains information recorded for each MAFX, like whether the MAFID is included in the ACS universe or its status as a group or living quarter. This enables the demo frame team to rank candidate MAFIDs accordingly, without having to store and maintain the 16, ~1TB MAFXs used to create it. The MAF\_Master’s efficiency is relevant for researchers too, its size is only ~120GB.

Having identifiers for addresses or people is not enough. Migration research requires datasets with these identifiers on them to be combined to make a person/place table that also records the time the record is seen. Then a time series for a person can be built from the various records showing a person’s moves through time. Key administrative datasets include: the Internal Revenue Service’s 1040 and 1099 data, Veterans Service Group of Illinois’ consumer referential database, the Social Security Office’s records, the National Change of Address Files, American Community Survey data, Decennial Census data, etc. Note that some of these datasets are from third parties, like the Veteran Service Group of Illinois’ consumer referential data.

METHOD

Across all the chapters, I will use demoframe products to analyze the intersection between U.S. people and places. In this first chapter, I will analyze the general people and place coverage between the most used, large-scale data sources and the demoframe products. In the second, I will estimate measures of migration with these sources to illustrate the validity of this approach for migration. The third chapter will apply these concurrently validated approaches to a disaster migration case study. I turn now to the first chapter and analyzing the person and place coverage between the large-scale datasets. I propose using the datasets available in a Federal Statistics Research Data Center (FSRDC), including the 2020 Decennial Census Edited File (CEF), The American Community Survey micro data (ACS), The Demographic Frame extracts, and all other datasets included in the Person Place Table (which is also available in an FSRDC) to make a business rules approach to person place matching. The Person Place Table includes information from nearly a 1,000 source vintages including the U.S. Census Bureau’s version of the Social Security Administration’s Numerical Identification System (CNUM), data from the U.S. Postal Service’s National Change of Address File, IRS data, and state aid program datasets including the Supplemental Nutrition Assistance Program, the Temporary Assistance for Needy Families, or WIC.

These datasets all identify people using PIKs, and addresses through MAFIDs. The dates of the datasets, or the dates on the records themselves, can be used to identify when a particular person is at a particular address.

There are two modeling approaches used here that utilize the Person Place Table as the main input: the demographic frame extracts, sometimes referred to as person place models (PPM), which make predictions through machine learning and statistical models for a given extract year, and the business rules approach to person place matching, which uses flexible logic for the assignment of person\place pairs. MAFID and PIK identifiers are never repeated and entirely unique.

Both methods which utilize the Person Place Table, the business rules approach to person/place matching and the demoframe extracts, obtain their universe (or sampling frame where everyone is selected) differently from the Decennial Census or the ACS. These data products have a frame of addresses, and these are selected or sampled. The demographic frame extract and the business rules approach both start with a master PIK list: a list of all PIKS ever verified. They then utilize records like the CNUM that indicate a death in the period, identifying those who have died, and excluding those born after the reference date. Those who die during the reference period are kept and marked with a mortality attrition code, which is important to differentiate for disaster-related migration.

*The Demographic Frame Extract*

The demoframe extract offers four machine learning models which identify the best PIK/MAFID pairs for a given year and reference date: an elastic net, random forest, logit, and boosted tree model. They are trained on the extract year’s ACS data as a truth set and then uses the sources from the past two years in the Person Place Table to create PIK/MAFID pairs. Other features used for training include the sourceid, or the characteristics of a particular source, and the date a source was considered valid. Each model seems to have different strengths and weaknesses and there are different versions of the demoframe extracts with various reference dates and coverages. The assignment process and logic behind the PIK/MAFID pairs is opaque for each model. However, broadly speaking, the models prefer PIK/MAFID pairs with many corroborating sources, with higher quality sources, and more recent sources. It also has a feature which prefers mafids that were considered valid housing units during the last decennial census.

Internal analyses suggest using the random forest and logistic regression models over the elastic net and boosted tree models (Demographic Frame Team 2025). They suggest using either the random forest or logistic models based on their internal analyses, and so because the projected probability for the PIK/MAFID pairs are unimportant for this project, I use the random forest model whenever the demoframe extracts are used.

The business rules approach has utility as an residency estimator when paired with the last whole-universe dataset. This approach basically uses business rules and sources to update the last whole-universe dataset from one year to another, but it can also be used to update a dataset at any interval or when the assumptions of the data no longer hold enough to update. Such could happen if coefficients from a machine learning dataset are used in a period significantly different from the training data’s. OLS and other traditional estimation methods would suffer from the same problem.

The business rules approach uses different estimation strategies based on the research goal. When the objective is to identify movers after a particular event, the BRAPPA will use sources after the reference date. When the objective is to update a particular data product with business rules, BRAPPA will use sources from before and after the reference date. Both versions of the BRAPPA will use the MAFID from the last whole-universe data source (usually a demoframe extract or the Decennial Census) for PIKs without corroborating sources in the mini–Person Place Table’s slice of the time series. This will find those PIKs who might have been missed in a strictly prospective method.

When these methods are problematic or we are trying to make predictions about the present or near future, the BRAPPA will utilize a retrospective method. This is less than ideal broadly speaking, but especially for those affected by some migration-inducing event.

*The Business Rules Approach to Person/Place Matching: A Novel Contribution*

The business rules approach, in a nutshell, will start with the PIKs that meet the age thresholds for a particular reference date, take records where respondents verify their addresses personally (and without incentive) as truth, use business rules to choose between potential MAFIDs, puts children under 16 at the same address as their parents, and then use the last entire-universe-source observed MAFID for the remaining PIKs. There are idiosyncratic decisions made in this baseline. I will compare a few prospective variants in the process outlined below to the CEF and ACS. I will outline these variants after the main synopsis.

I detail each step below.

After assembling the master PIK table, the business rules approach will then select records from the Person Place Table from a year around the reference date. This makes a mini–Person Place Table and prepares for efficient searching because the whole Person Place Table is quite large. It then uses the National Change of Address File provided by the U.S. Postal Service to identify those who have moved temporarily within a month of the reference date, and it assigned the PIKs that show up in this interval the respective MAFID. A temporary move can last from one month to six and is updated in each monthly vintage of the National Change of Address File. I will then assign MAFIDS to PIKs who indicate a permanent move within 3 months of the reference date. Permanent moves are retained on the National Change of Address File for about a year.

Some movers move out of the country and the National Change of Address file records these moves. These PIKs are marked with a foreign move attrition code. Some movers are missing there the moved to, but we do know where they moved from. I keep this information as a ‘not\_mafid’ and ensure subsequent matching does not use this MAFID.

PIKs that participated in the ACS with an interview date within a month of the reference date will be assigned to the MAFID they had at the time of the interview.

\*\*\* THIS PART IS VERY SUBJECT TO CHANGE (I got a lot of ideas that get better and better)\*\*\* From here the mini-Person Place table will be collapsed with a “GROUP BY” function, that will count the number of sources supporting a particular PIK/MAFID pair, with the first observed date for that pair being recorded. PIK/MAFID pairs will be selected first if the first observed date is within a month of the reference date, the number of sources corroborating the PIK/MAFID pair is four or more. When there is no available match for those conditions, PIK/MAFID pairs will be chosen if the difference between the first observation date and the reference date is less than 90 days, and the number of corroborating sources is larger than two. Failing those conditions, the PIK/MAFID pairs with the highest source count will be pick with the earliest first observation date being the tie breaker.

The relations table is a table that identifies the mothers and fathers for each PIK. It uses information from all the vintages of the Census Household Composition Key to identify the parents. The relations table will be used to put PIKs who do not make records, usually young children, with their listed mothers, and then fathers, in that order if one is missing.

For all others who have not been assigned a MAFID at this point, I will use the MAFID assigned at the last whole universe data product. This is usually the most current version of the demo frame extract but could be the Decennial Census.

Using a table that collates all the MAFID information for each from 2010 to the present, I can make extracts for any of these periods. This combined table of MAFIDs is important because geography changes frequently and addresses in a particular zip code or county can be moved to other localities. Buildings represented by MAFIDs also change their purpose from time to time and assigning people to a MAFID that used to be an apartment but is now a business with no live-in residents is unacceptable. As the geography division of the Census Bureau retires or combines duplicate MAFID’s, this table tracks which MAFIDs are active, and the geography of each MAFID at any given time. Using the Maf\_Master table, I change retired MAFIDs to the designated ‘surviving MAFIDs’ of that year. Many of the MAFIDs identified in the Person Place Table are MAFIDs that represented the same address at different points in time. The transitive nature of MAFIDs and addresses necessitates a longitudinal approach like the Maf\_Master.

At the end of this process, we have a table of PIKs alive close to the reference date with markers for those who move outside of the country or die within a user-specified interval from the reference date, every PIK is assigned to a MAFID, and demographic information, such as race, ethnicity, or sex can be joined from the last whole universe data product.

*Assessing the Comparability of BRAPPA, Decennial Census, and Demoframe Extract Matching*

It is common to compare new data products with older, assumed valid, data products to assess the performance and validity of the new data product, argument of concurrent validity. We can additionally simulate the predictive validity of the BRAPPA by utilizing the retrospective strategy with sources published before April of 2020. Evaluation for predictive validity is usually the same as concurrent validity: correlations at or above .7 are usually considered sufficient (Streiner, Norman, and Cairney 2015), though these rules vary among researchers. Pearson correlations are for continuous data, and the BRAPPA pik/mafid pairs are nominal and so tetrachoric correlations (framed as dichotomous match/not-match) are the preferred approach.

To compare the performance of these various methods, I will analyze the percent match between the 2020 Decennial Census, the Demoframe 2020v3 extract, all BRAPPA variants ( including the combined retrospective and prospective strategies and the retrospective only strategy). I will analyze the percent match of the data products with each other. Where computationally possible, I will calculate Cohen’s Kappa, a common measure of inter-rater reliability, between the various data frames. These will build an argument that the BRAPPA and the PPM can be used in cross sectional analysis for people and places and are similar to the quality of a decennial Census. The PPM already has analyses revealing it is close the Decennial census in a lot of ways, but this looks at PIK and MAFID converage and compares these with the ACS as well.

Table shells of these analyses and comparisons are in the results section.

RESULTS

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| TABLE 1. Characteristics of 2020 Data Products. | | | | | |
| Data Product | Reference Date | Total Included PIKs | Total Included MAFIDs | PIK Coverage | MAFID Coverage |
| 2020 Decennial Census Edited File (CEF) |  |  |  |  |  |
| American Community Survey |  |  |  |  |  |
| BRAPPA\_full |  |  |  |  |  |
| BRAPPA\_full\_prospective |  |  |  |  |  |
| BRAPPA\_full\_retrospective |  |  |  |  |  |
| RF PPM 2019 |  |  |  |  |  |
| RF PPM 2020 |  |  |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| TABLE 2. Characteristics of 2020 Data Products By Moving Status. | | | | | |
|  | | For Movers | | For Non-Movers | |
| Data Product | Reference Date | PIK Coverage | MAFID Coverage | PIK Coverage | MAFID Coverage |
| 2020 Decennial Census Edited File (CEF) | April 1 2020 |  |  |  |  |
| American Community Survey | Anytime April 2020 |  |  |  |  |
| BRAPPA\_full | April 1 2020 |  |  |  |  |
| BRAPPA\_full\_prospective | April 1 2020 |  |  |  |  |
| BRAPPA\_full\_retrospective | April 1 2020 |  |  |  |  |
| RF PPM 2019 | April 1 2020 |  |  |  |  |
| RF PPM 2020 | April 1 2020 |  |  |  |  |

*Comparisons between 2020 Data Products*

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TABLE 3. Comparisons Between 2020 Data Products and the CEF | | | | | | |
|  | For Movers | | | For Non-Movers | | |
| Data Product | %Agreement | Correlation | Kappa | %Agreement | Correlation | Kappa |
| American Community Survey |  |  |  |  |  |  |
| BRAPPA |  |  |  |  |  |  |
| BRAPPA\_prospective |  |  |  |  |  |  |
| BRAPP\_retrospective |  |  |  |  |  |  |
| RF PPM 2019 |  |  |  |  |  |  |
| RF PPM 2020 |  |  |  |  |  |  |

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| TABLE 4. Comparisons Between 2020 Data Products and the ACS | | | | | | | |
|  | For Movers | | | For Non-Movers | | | |
| Data Product | %Agreement | Correlation | Kappa | %Agreement | Correlation | Kappa |
| 2020 Decennial Census |  |  |  |  |  |  |
| BRAPPA |  |  |  |  |  |  |
| BRAPPA\_prospective |  |  |  |  |  |  |
| BRAPPA\_retrospective |  |  |  |  |  |  |
| RF PPM 2019 |  |  |  |  |  |  | |
| RF PPM 2020 |  |  |  |  |  |  | |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TABLE 5. Comparisons Between 2020 Data Products and the BRAPPA | | | | | | |
|  | For Movers | | | For Non-Movers | | |
| Data Product | %Agreement | Correlation | Kappa | %Agreement | Correlation | Kappa |
| 2020 Decennial Census Edited File (CEF) |  |  |  |  |  |  |
| American Community Survey |  |  |  |  |  |  |
| RF PPM 2019 |  |  |  |  |  |  |
| RF PPM 2020 |  |  |  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TABLE 6. Comparisons Between 2020 Data Products and the 2020 Random Forest PPM | | | | | | |
|  | For Movers | | | For Non-Movers | | |
| Data Product | %Agreement | Correlation | Kappa | %Agreement | Correlation | Kappa |
| 2020 Decennial Census Edited File (CEF) |  |  |  |  |  |  |
| American Community Survey |  |  |  |  |  |  |
| BRAPPA |  |  |  |  |  |  |
| BRAPPA\_prospective |  |  |  |  |  |  |
| BRAPPA\_retrospective |  |  |  |  |  |  |
| RF PPM 2019 |  |  |  |  |  |  |
| RF PPM 2020 |  |  |  |  |  |  |

CONCLUSION

The PPM and BRAPPA are novel contributions to address the lack of coverage or lack of data available in surveys. The PPM (and hopefully BRAPPA) are better performing than even the best 3rd party dataset reviewed by the Pew Research Center (Chun et al. 2021; Ortman and Knapp 2023) (This is NOT a disclosure violation, the PPM models are validated and their performance is published). This performance indicates they adequately represent a population at a given point in time, but can they measure migration?

Chapter II: Estimating Migration

The purpose of this chapter is to connect the first and final chapters. We explored the accuracy of PIK and MAFID assignment in chapter I. In chapter III, I will apply these methods to a hurricane case study. How accurate are migration estimates derived from the demographic frame versus the traditional approaches? I’ll again use the ACS and Decennial Census as benchmarks to argue for the applicability of the BRAPPA and PPM. Then I will calculate various network measures for the migration matrices and model them in a unique form of multiple regression, the multiple regression with quadratic assignment procedure (MRQAP). This chapter starts with the theories highlighting migration processes, which then inform our modeling and estimation strategies.

While many disciplines have been theorizing about migration for a long time, U.S. sociology begins in the 1940’s (Bijak 2006) with Stouffer’s (1940) intervening opportunities. He posits that migration to a place will increase as the number of opportunities (especially jobs) available at a place increases. The likelihood of migration will decrease relative to the number of places and the number of opportunities available at closer places. This first sociological step identifies the importance of opportunities, as well as establishes a preference for less distance.

Lee (1966) builds upon the idea of intervening opportunities by adding push factors, or the factors that could push a person out of an origin. This literature then further divides push factors into hard and soft push factors. High interest rates, poor schools, etc. are soft push factors compared to hard push factors: war, natural disasters, or humanitarian crises. Disasters are hard push factors. Most of the literature on disaster migration acknowledges the primacy of push factors in life threatening situations, but recent work has identified trends in internal migration related to decline in precipitation or changes in temperature (Berlemann and Steinhardt 2017).

Taylor (1984) notes that migrant social networks seem to be very important pull factors. Ties in a destination diminish the cost of moving by allowing migrants access to social support, information about a place, and the capacity for more preparation by ties already at the potential destination.

Trans-national spaces (Bilecen, Gamper, and Lubbers 2018; Faist 2015; Roth 2009)builds upon this to conceptualize a meso-level space where migrants negotiate their identities between places. Trans-national spaces are the social capital networks and institutions that bridge places together and help in-group members through the transition wholistically. These are places facilitating integration, while establishing a separate identity from either origin or destination identities.

Interdisciplinary theories with sociology exist as well. Institutional theory (Massey et al. 1993) compliments the network-based insights of Taylor’s observation on migrant social networks by examining the connections that migrants have with institutions, like NGOs, corporate recruiters, counselors, and even irregular institutions like human smuggling or trafficking. The emphasis on institutions dovetails into institutional theory of economics well, creating a de facto hybrid, cross-disciplinary theory.

Relatedly, Cumulative Causation is a theory put forward by Massey (Fussell and Massey 2004; Massey 1990).It asserts that migration is an evolutionary process that changes the origin and destination. The people involved undergo a transformation from migration too, returning with more human and social capital (not to mention the other benefits like income). Migration will redistribute the land and other capital in a sending place as well, and these incentives can instigate a migrant culture, where migration is romanticized for its capacity for capital gain, and the costs of migrating can be reduced with a strong migration stream (as pointed out by Taylor’s migrant networks or the trans-national spaces literature) and institutions at the sending and receiving points of the stream. These externalities to migration can reinforce the migration process such that migration takes on a macro-level stream as opposed to many individual actors making many unique individual decisions. Evaluations of this theory have found cumulative causation has a lot of explanatory power for rural and smaller communities, but less predictive power for urban or larger communities. In other words, the migration processes may depend on who and where a migrant is going (Fussell and Massey 2004).

Economic theories of migration also exist and have been influential in sociological theories, resulting in some of the hybrid theories discussed above.

Migration in the neo-classical tradition is a disequilibrium phenomenon where capitalistic economies with a surplus of labor will give workers to economies with a surplus of capital. Capital movements and labor movements go in both directions and migration of these factors will cease once equilibrium is reached.

The micro economic version of this is that individuals are motivated to increase their lifetime earnings. Because of this motivation, workers should permanently move to wherever seems to have the best return to lifetime earnings, with a penalty imposed per distance of the opportunity.

Neoclassical economics does not describe return migration, nor the tendency of humans to organize their economic outputs in collective households (neoclassical economics assumes individuals are all motivated by their own lifetime earnings). There are also migration flows without wage differentials that are unexplained by neoclassical economics.

The new economic theory of migration is a micro economic theory revolving around households as the unit of analysis. These households are incentivized to mitigate risk, not maximize their earnings. When the source of risk in the sending context is addressed or the life cycle of the household has rendered a previous untenable risk tenable, this theory expects the return migration of the household, which is a great expansion on neoclassical economics described above.

Dual labor market theory describes the incentives for migration at a destination. The labor market is divided into two labor markets. There is a capital-intensive market and demand in this market is stable (not stationary). Workers in this market are usually skilled, and disruptions in this market are rarer than in the other market. There is also a labor-intensive market, which handles a lot of variant demand. This labor market is full of low-skill workers whose jobs are unstable. No one really wants to be in the labor-intensive market, but firms span both markets and need people in the labor-intensive market. There are two strategies firms could use to incentivize workers to work in the labor-intensive market.

First, they can increase compensation for labor-intensive workers. This strategy can result in wages increasing all through the hierarchy as workers observe a group is getting wage increases and apply pressure for their own wage increases. The second option is more popular: Obtaining workers from another place to work for low wages. This saves money for the firm. Because there are no other options to obtain labor, companies lobby the government for more migrants and for fewer obligations for their foreign workforces.

There is also world systems theory, which is hybridized with economic ideas. World systems theory is about the processes affecting the sending of migrants. As capitalism/modernity progresses, markets transition from an agrarian or industrial economy to a service economy. These advances take place in the world “core” or the developed countries usually in the global North, and “periphery” and “semi-periphery” regions. A flow of goods and capital from core to periphery regions is counter balanced by a reverse flow of labor to periphery countries. In core regions, manufacturing jobs become less and less desirable and demand for these jobs increases, creating an opportunity for migration. In periphery regions, the increased production from technological advancements or capital investments results in less demand for workers. These workers are uprooted by these circumstances and incentivized into low paying, labor intensive positions in the core. There are many links from core countries to periphery countries beyond economics, the cultural, historical, linguistic, etc. factors are important too, which separates this from purely economic theories.

As noted by Massey et al. (1993: 448), in the world systems approach “international migration ultimately has little to do with wage rates or employment differentials between countries; it follows from the dynamics of market creation and the structure of global economy”. Special attention is paid to the asymmetric relationship between colonies and colonizer historical relationships; former colonizers being seen as having an advantage in trade. This is controversial, because free trade is seen as reducing income and employment disparities, and thus also migration. This theory is not elucidated mathematically and so is difficult to use in predicting future migration. Its emphasis on the global system and interconnected nature of place and people compliments a rising paradigm of examining the system(s) of migration, analyzing sending and receiving geographies at the same time.

To summarize some key takeaways from the various theories of migration: Migration is inherently about opportunities and consequences. Opportunities to avoid death, disease, disaster, and risk are reasons to move, or push-factors. Opportunities to gain money, security, be with family and friends can be reasons to move or to stay. There is a demand in receiving countries, which usually are more advanced service-based economies and often have a history of exploiting the resources and people from sending countries, for cheaper labor. There is a supply of migrants from sending countries who often are looking to increase their incomes or mitigate risk/overcome a challenge in their community. The interconnected relationship between sending and receiving places suggests a wholistic approach: analyzing the matrix of sending and receiving places at once. The systems and covariates of migration can cause migration streams themselves through cumulative migration.

Because every migration is a zero-sum event, origins and destinations are frequently theorized and analyzed together. One way to consider the origins and destinations together and handle the aggregation required to respect federal data standards is to create matrices of migration from these aggregations (Curtis, Fussell, and DeWaard 2015; Hauer, Holloway, and Oda 2020; Johnson, Bland, and Coleman 2008). These matrices usually combine the immigrants and emigrants by column and row, with cell counts particular to a specific place, i.e. the net migrants from placei to placej in columni, rowj. Researchers then analyze the migration system rather than individuals who migrate or migration’s effects on a single geography.

Currently the U.S. Census Bureau releases migration estimates using the American Community Survey. Estimates from the ACS combine serval years of data for smaller areas because of the ACS’s rotating sampling schedule. This means that migration matrixes are only available in five year blocks, the 2016-2020 block being the most recent (United States Census Bureau 2025d). This aggregation of time periods is a severe limitation for analysis.

*Network Analysis Insights for Migration*

Networks are a analytical tool for interlinked phenomena. Migration has many inter-related domains and while the social network components of migration are increasingly attended to, the physical locations are also linked by recurrent flows. These spatial relationships are less understood but still explored, the simplest models are called gravity models.

Migration gravity models include a term for the ‘pull’ a location has, or the tendency to attract migration, often the population size, or expected income of jobs at that location (Borjas 1994; Windzio 2018). This term interacts with a term for the distance between locations, which together represent the distance decay of the attractiveness for a place. In other words, some places are more attractive than other places, but the attractiveness of locations depends on how close the location is. As predicted under the neoclassical economics of migration, a location that is too far will have too much cost and or work involved to be a practical option, and we can model this with a variable to represent the value of moving to a particular place multiplied by a decay for the distance (Borjas 1990; Massey et al. 1993).

Gravity models can be considered a unique case of network analysis, where the distance decay rate between two places can describe the tie between two nodes. A tie is any relationship between two nodes, or two units of analysis, that can have relationships between other nodes (Perry, Pescosolido, and Borgatti 2018). The nodes are often people and the ties the relationships between them, but it can also be used to analyze locations as nodes and the relationships between them (like distance, money and trade flows, or tourism and migration flow) as ties (Danchev and Porter 2020, 2021; Windzio 2018).

While the application of network analysis to migration networks is fairly recent (Windzio 2018), there are many network analysis measure that represent ideas from the migration literature.

Degree is the number of connections a node has and has the flexibility to be directional when practical (Robins 2015). In-degree is the number of connections coming into a node, like migrations to a particular city, while out-degree is the number of connections leaving. The traditional theories of migration point out a tendency for some locations to be the destination for many migration streams, sometimes called hubs (Bell, Charles-Edwards, Kupiszewska, et al. 2015; Newman 2018). Degree puts the focus of the analysis on the tendency to have connections with other places and describes the average effect of these relationships between location rather than the characteristics of the location (Danchev and Porter 2020). Thus, people from places with a large number of in-degree could represent popularity while people from places with many out-degree connections may have many accessible options for migration. Additionally, we can consider the degree of the entire network as a whole, the number of connections between all the nodes, which allows for comparisons across different matrices, either for different places, or times with different conditions.

Counter-migration streams, or a current of migrants returning from places that are usually destinations to places that are usually origins, has been a frequent and time-honored observation by theorists of migration (Grigg 1977:112; Ravenstein 1885:199). Migrants may be returning to areas after obtaining resources (Stark and Bloom 1985; Stark and Taylor 1991). In network analysis, this is represented by reciprocity or the tendency for nodes, counties or regions, to both have in/out edges with each other (McCarty et al. 2019; Perry et al. 2018).

Closure or clustering of migration is described in migration theory. Many locations are stepping-stones to other locations(Bell, Charles-Edwards, Kupiszewska, et al. 2015; Bell, Charles-Edwards, Ueffing, et al. 2015; Bernard, Bell, and Charles-Edwards 2014) and a particular so-called destination may just the point of departure after a little while. The concept of transitivity, or in migration, how transitory a place is compared to other places in the matrix, can be analyzed. The Geometrically Weighted Edgewise Shared Partners (GWESP) is a measure used to analyze the transitivity of the networks. It is simply a count of transitive edges, triangles if we are looking at a map of nodes with lines connecting their edges, while weighting the number of edges each node has to ensure equal representation of edges regardless of the degree of the node. This is an important part of estimating network models because without a term like GWESP, the parameter space “corresponds to nearly degenerate distributions” (Snijders et al. 2006:1). The term also helps separate out the effects of similar but distinct ideas like homophily and reciprocity (Hunter and Handcock 2006).

Homophily is the tendency of nodes to create edges with nodes with similar attributes, and is a ubiquitous part of human social networks (Bratter and O’Connell 2017; Choudry, Williams, and Black 2017; Portes 1998; Windzio 2018). Cumulative causation and trans-national spaces address identify the tendency for migrants from origin countries to cause additional migration from the origin country through interpersonal connections. At an aggregate level, we would expect individuals to choose to resettle in places with many other migrants from their country of origin. This tendency is captured under a homophily term in network analysis.

Chain migration, or the tendency to migrate in with multiple destinations before arriving at a final destination (Lucas 1997). These reflect relationships between aggregates and, when combined with cumulative causation theory, these relationships can persist; as seen qualitatively (Fussell and Massey 2004; Massey 1990). These chains are difficult to identify at a individual level but should be identifiable at a macro level. Network analysis can observe the clustering of edges for localities in the migration matrix. The strength of the community can be captured in a modularity measure called Q, which examines the number of edges among a group compared to the expected number of edges for the rest of the graph (Borgatti et al. 2022; Robins 2015). Then various algorithms can be used to identify groups maximizing Q for the various groups, WHAT PACKAGES DO I HAVE ACCESS TO IN EC2?

The most important insight from network analysis to studies of migration is perhaps the recognition that neither the nodes (the locations) nor the edges between nodes (the streams between locations) are independent phenomena. The lack of independence violates the assumptions of many inferential tools, including ordinary least squares regression. There have been several innovations in response to this, including the multiple regression with Quadratic Assignment Procedure (MRQAP) (Krackhardt 1988). There are also Exponential Random Graph Models (ERGMs), which are generally better and we will use in chapter three, but ERGMs are harder to estimate and often need to use simple models for large networks, so MRQAPs are often necessary. The most popular MRQAP used today is the double semi-partialing procedure (David Dekker, David Krackhardt, and Tom Snijders 2007), which I lay out here:

Consider a basic linear model for square matrix data like

graphic file with name M1.gif

where *Y*, *X*, and *E* are *n* × *n* matrices, *β* is a scalar, *Z* is an *n* × *n* × *q* array, and *γ* is *q* × 1. The diagonals of the matrices are always ignored. The null hypothesis is

graphic file with name M2.gif

And the relationship between X and Z assumed to be not independent and assumed to be

graphic file with name M3.gif

Where *V* is an *n* × *n* matrix. The situation *δ* ≠ 0 is called collinearity. The nonparametric approach to square matrix data means here that the residuals associated with the *n* objects are exchangeable or, in other words, the matrices *E* and *V* are invariant under permutations of rows and columns simultaneously by the same permutation. Whenever the term *permutation* is used, it will be assumed that this permutation acts on rows and columns simultaneously, which is described by the π symbol.

The residual between X and Z can be defined as

graphic file with name M22.gif

, where Inline graphicis the OLS estimate for the model

|  |
| --- |
| graphic file with name M24.gif |

Row/column permutations are designated by π, and the residuals are permuted. The resulting permutations are isomorphic to the original, so all features of the original matrix are retained except for relationships about the order of the objects. Repeating these permutations can generate a reference distribution for values under the set of all permutations which can then be used to test the null hypothesis that two matrices are independent. The model

|  |
| --- |
| graphic file with name M25.gif |

is used to obtain reference values for the test statistic. The rationale being that under the null hypothesis *β* = 0, the above model for *Y* is the same as the original model

graphic file with name M1.gif

, and if the estimation error Inline graphicis negligible, the permutational invariance assumption for *V* implies that

|  |
| --- |
| graphic file with name M27.gif |

has the same distribution as *V*.

METHOD

I propose creating net migration matrices for every county/county-equivalent in the United States by subtracting the number of movers to a county by the number of movers from that same county, repeated for every county. I use the best BRAPPA model, the PPM’s logistic regression model, and the 2020 Decennial Census to identify where people are in 2020. I subtract each data source from the PPM logistic regression model for 2019. This matrix operation uses algebraic logic: because each data product is subtracted from the same data, the PPM from 2019, comparisons between the difference matrices reveal only the dataset’s characteristics and not the PPM from 2019. This analyzes the differences between migration matrices built from the different methods. While taking the results of this method does assume that the 2019 PPM is accurate, the purpose of this analysis is not to present any migration matrices as true, but rather to compare the matrices that could be made from the respected 2020 Decennial Census with other methods if the 2020 Decennial Census could be used to create yearly migration matrices.

The validity of the BRAPPA and PPM are argued similarly to the first chapter, we compare the matrices based on the different approaches, assuming the decennial census represents the ground truth. However, I also add modeling and an MRQAP to be able to estimate the marginal effects between the various estimates.

RESULTS

(NOTE for Dr. Amaral: I was making these analyses with synthetic data before the shutdown. The tables and maps will be fleshed out once I know what is possible and what the variables can look like)

|  |  |  |
| --- | --- | --- |
| TABLE 7. Correlations and Distance errors between PPM 2020, BRAPPA, and CEF Migration Matrices (as ordered vectors) | | |
| Model | Comparison Against CEF 2020 | |
|  | Correlation | Average Absolute Difference |
| BRAPPA | NA correlation | 0 miles |
| PPM2020 | NA correlation | 0 mi |
| Note: average distance difference calculated with population centered meanoids for each county, where addresses in the same county are given a distance of 0. The meanoids represent the population centers in 2020 (United States Census Bureau 2020). | | |

FIGURE 1. Maps comparing PPM 2020, BRAPPA, and CEF Migration Matrices

Facet-Wrap 3 gradient choropleth maps here

FIGURE 2. Hotspot Analyses Map

3 facet-wrap with significant hot/cold spots highlighted

FIGURE 3. Fruchterman and Reingold Visualization of BRAPPA Migration Matrix With Region and Cluster Algorithm Colors: Maybe a Facet Wrap to show both.

TABLE 8. MRQAP Of County-level Charateristics Impact on Migration.

RESULTS

CONCLUSION

The comparisons between the BRAPPA and PPM2020 should reflect the results in chapter one: that the BRAPPA is essentially a way to update a PPM for a particular period based on the sources in the administrative record. We observed some of the characteristics of the migration matrix in the United States and found unique insights from a network analysis approach: we obtained estimates on transitivisty and identified clusters of counties that send many people to each other. We identified the impact that having a higher gross domestic product has on county level migration. This will lay a good foundation for a case study with no available comparisons, the migration matrix during hurricane Ian.

Chapter III: A Hurricane-Migration Case Study

Disaster based migration typically situates itself under other migration theories. Typical migration theory processes are applied with careful consideration to the context of the disaster, often framed and treated as a push factor (Curtis et al. 2015; Hauer et al. 2020; Zhou et al. 2022).

However disaster migration also has its own unique emphases. Berlemann and Steinhardt (2017) identify geography and climate as central push and pull (or stay) factors for disaster migration because the climate, geography, and characteristics of the geography are what brought people to the location in the first place. While climate and geography are natural push and pull factors, many feel that isolating the push and pull effects of climate or geography is very complicated, perhaps impossible, because of how climate and geography interact and endogenously relate with other factors like economics, social networks, health, food, politics, and policy (Piguet, Pécoud, and de Guchteneire 2011). To illustrate the circuitous nature of these relationships: a draught in South America often leads to migration to the United States, but for Mali less rain leads to lower levels of migration (especially to other African countries and France) because of policies that tighten credit constraints and consequently raise food prices. The interconnected nature of variables seems to recommend an approach that can analyze many types of variables simultaneously while considering the continuous nature of time, like an Exponential Random Graph Model (ERGM).

*ERGMs in Migration*

ERGMs present partial derivative coefficients, just like an ordinary linear regression. I include measure of county-level GDP, racial proportions, spatial distance from centroids, \*\*\* NEEDS FLESHING OUT WHEN KNOW WHAT IS COMPUTATIONALLY POSSIBLE

Similarly to MRQAPS, ERGMs are not often used to measure migration. Though some examples do exist (Windzio 2018), they are often used at a country-level where the number of nodes is just under 250. ERGMs are computationally intensive and modeling an ERGM with thousands of nodes is a careful balance of only including theoretically relevant terms, avoiding adding too many terms to feasible compute, and avoiding terms that will bring the parameter space to correspond with nearly degenerate distributions, where the probability is heaped up to only a couple of terms and poorly fit the empirical data (Snijders et al. 2006).

Memory is the tendency for past migration patterns to continue in the present. Lagged variables are generally an important part of any time series modeling and the justification for including lagged variables in the model is most clearly identified under cumulative causation theory, where past migration streams form the connections and resources for future migration (Bachmeier 2013; Massey 1990). There is a notable caveat to cumulative causation that we should see in a migration matrix, cumulative causation seems to apply best to rural, medium-sized locations (Fussell and Massey 2004). In the context of Hurricane Ian, the migration matrix from 2022 can be used to see how the disaster-affected locations change (or do not change) their typical migration patterns.

*Identifying Areas Affected by Disasters*

In U.S. disaster migration work, it is common to analyze counties where Federal Emergency Management Agency issues an emergency declaration (Curtis et al. 2015; Johnson et al. 2008). These counties are sometimes analyzed against counties without an emergency declaration. Recent work has begun to consider the entire matrix of migration relationships: the ties each county has with each other county in send and receiving migrants (Curtis et al. 2015; Hauer et al. 2020).

^OTHER RESEARCHERS ^

*Hurricane Ian*

METHOD

I create a net migration matrix and analyze the flows with an ERGM. I will apply and extend the methods used in the previous two chapters on the Hurricane Ian case. I will estimate an Exponential Random Graph Model (ERGM) to model the migration patterns of the United States while controlling for other factors and accounting for the non-independence between dyads. These are seldom done in studies of migration (Windzio 2018). ERGMs are social network models designed to control for modeled effects to present partial derivative coefficients and account for the non-independence of dyads when node pairing can depend on the actions of at least one person (Fawcett 1989; Windzio 2018).

I calculate the network measures discussed in chapter two: gwesp, homophily, centrality, etc. .

RESULTS

CONCLUSION

I propose an ambitious dissertation: to create a new way to estimate migration using administrative records. I will compare and validate estimates from various respected sources to build the case for using a business rules approach to matching people to places in specific time periods. I finally propose taking the validated business rules approach to a hurricane case study to examine how disasters affect the internal migration matrices of the United States.

There are, of course, limitations to this work. ERGMs are difficult to estimate, and these can limit their potential. MRQAPs are more flexible but do not account for the endogeneity that ERGMs can. While approaches using administrative records cover most people, there will usually be some demographics underrepresented in the data. However, these limitations are the same limitations we have always had. No modeling strategy is perfect. Our surveys are already inheriting the coverage issues of administrative data because the sampling frames are administrative data like the MAFX or a PPM.

This new approach addresses many problems with older approaches: the fact that places are not independent observations and typical approaches do not account for this, that once venerable surveys are wanning inaccurate (Hyatt et al. 2018), and the patchwork coverage of migration data sources. It is feasible to compute on common computers and yields estimates that can be tweaked with scientific theory, and in this sense is easier to work with than black box methods.

REFERENCES

Alvaredo, Facundo, Anthony B. Atkinson, Thomas Piketty, and Emmanuel Saez. 2013. “The Top 1 Percent in International and Historical Perspective.” *The Journal of Economic Perspectives* 27(3):3–20. https://www.jstor.org/stable/41955542.

Ambel, Alemayehu, Kevin McGee, and Asmelash Tsegay. 2021. *Reducing Bias in Phone Survey Samples: Effectiveness of Reweighting Techniques Using Face-to-Face Surveys as Frames in Four African Countries*. Policy Research Working Papers. The World Bank.

Anthony Wray, University of Southern Denmark, United States Federal Security Agency Social Security Board, and United States Social Security Administration. 2024. “Numerical Identification Files (NUMIDENT), 1936–2007.” Inter-university Consortium for Political and Social Research (ICPSR). doi:10.3886/E207202V1.

Baines, Rebecca, Sebastian Stevens, Daniela Austin, Krithika Anil, Hannah Bradwell, Leonie Cooper, Inocencio Daniel Maramba, Arunangsu Chatterjee, and Simon Leigh. 2024. “Patient and Public Willingness to Share Personal Health Data for Third-Party or Secondary Uses: Systematic Review.” *Journal of Medical Internet Research* 26(1):e50421. doi:10.2196/50421.

Berlemann, Michael, and Max Friedrich Steinhardt. 2017. “Climate Change, Natural Disasters, and Migration—a Survey of the Empirical Evidence.” *CESifo Economic Studies* 63(4):353–85. doi:10.1093/cesifo/ifx019.

Bijak, Jakub. 2006. “Forecasting International Migration: Selected Theories, Models, and Methods.” http://www.cefmr.pan.pl/docs/cefmr\_wp\_2006-04.pdf.

Bilecen, Başak, Markus Gamper, and Miranda J. Lubbers. 2018. “The Missing Link: Social Network Analysis in Migration and Transnationalism.” *Social Networks* 53:1–3. doi:10.1016/j.socnet.2017.07.001.

Böhme, Marcus H., André Gröger, and Tobias Stöhr. 2020. “Searching for a Better Life: Predicting International Migration with Online Search Keywords.” *Journal of Development Economics* 142:102347. doi:10.1016/j.jdeveco.2019.04.002.

Boyle, John, Lewis Berman, James Dayton, Ronaldo Iachan, Matt Jans, and Randy ZuWallack. 2021. “Physical Measures and Biomarker Collection in Health Surveys: Propensity to Participate.” *Research in Social and Administrative Pharmacy* 17(5):921–29. doi:10.1016/j.sapharm.2020.07.025.

Bronson, Nathan, Thomas Lento, and Janet L. Wiener. 2015. “Open Data Challenges at Facebook.” Pp. 1516–19 in *2015 IEEE 31st International Conference on Data Engineering*.

Bureau of Labor Statistics. 2025. “CPS Response Rates.” https://www.bls.gov/cps/methods/response\_rates.htm.

Buzzelli, Michael. 2020. “Modifiable Areal Unit Problem.” *International Encyclopedia of Human Geography* 169–73. doi:10.1016/B978-0-08-102295-5.10406-8.

Call, Kathleen Thiede, Michael Davern, Michel Boudreaux, Pamela Jo Johnson, and Justine Nelson. 2011. “Bias in Telephone Surveys That Do Not Sample Cell Phones: Uses and Limits of Poststratification Adjustments.” *Medical Care* 49(4):355. doi:10.1097/MLR.0b013e3182028ac7.

Card, David, Raj Chetty, Martin S. Feldstein, and Emmanuel Saez. 2010. “Expanding Access to Administrative Data for Research in the United States.”

Chun, Asaph Young, Michael D. Larsen, Gabriele Durrant, and Jerome P. Reiter, eds. 2021. *Administrative Records for Survey Methodology*. Hoboken, NJ: Wiley.

Craig, Jean B., Joan M. Culley, Jane Richter, Erik R. Svendsen, and Sara Donevant. 2018. “Data Capture and Analysis of Signs and Symptoms in a Chemically Exposed Population.” *Journal of Informatics Nursing* 3(3):10–15.

Craig, Jean B., Joan M. Culley, Abbas Tavakoli, and Erik R. Svendsen. 2013. “Gleaning Data From Disaster: A Hospital-Based Data Mining Method To Studying All-Hazard Triage After A Chemical Disaster.” *American Journal of Disaster Medicine* 8(2):97–111. doi:10.5055/ajdm.2013.0116.

Curtis, Katherine, Elizabeth Fussell, and Jack DeWaard. 2015. “Recovery Migration after Hurricanes Katrina and Rita: Spatial Concentration and Intensification in the Migration System.” *Demography* 52(4):1269–93. doi:10.1007/s13524-015-0400-7.

Demographic Frame Team. 2025. “Demographic Frame Extract Notes- 2023v1.”

Devine, Jason, Spader Jonathan, and King Ryan. 2021. “2020 Census Data Review.” https://www.census.gov/newsroom/blogs/random-samplings/2021/04/2020-census-data-review.html.

Faist, Thomas. 2015. “Transnational Social Spaces.” *Ethnic and Racial Studies* 38(13):2271–74. doi:10.1080/01419870.2015.1058502.

Fawcett, James T. 1989. “Networks, Linkages, and Migration Systems.” *International Migration Review* 23(3):671–80. doi:10.1177/019791838902300314.

Feliciano, Cynthia, and Yader R. Lanuza. 2017. “An Immigrant Paradox? Contextual Attainment and Intergenerational Educational Mobility.” *American Sociological Review* 82(1):211–41. doi:10.1177/0003122416684777.

Fussell, Elizabeth, and Douglas S. Massey. 2004. “The Limits to Cumulative Causation: International Migration from Mexican Urban Areas.” *Demography* 41(1):151–71. https://www.jstor.org/stable/1515217.

Garson, George David. 2019. *Multilevel Modeling: Applications in STATA®, IBM® SPSS®, SAS®, R, & HLMTM*. 1st edition. Los Angeles: SAGE Publications, Inc.

Gourlay, Sydney, Talip Kilic, Antonio Martuscelli, Philip Wollburg, and Alberto Zezza. 2021. “Viewpoint: High-Frequency Phone Surveys on COVID-19: Good Practices, Open Questions.” *Food Policy* 105:102153. doi:10.1016/j.foodpol.2021.102153.

Groves, Robert M. 2011. “Three Eras of Survey Research.” *Public Opinion Quarterly* 75(5):861–71. doi:10.1093/poq/nfr057.

Harron, Katie, Chris Dibben, James Boyd, Anders Hjern, Mahmoud Azimaee, Mauricio L. Barreto, and Harvey Goldstein. 2017. “Challenges in Administrative Data Linkage for Research.” *Big Data & Society* 4(2):2053951717745678. doi:10.1177/2053951717745678.

Hauer, Mathew, and James Byars. 2019. “IRS County-to-County Migration Data, 1990‒2010.” *Demographic Research* 40:1153–66. https://www.demographic-research.org/articles/volume/40/40.

Hauer, Mathew E., Steven R. Holloway, and Takashi Oda. 2020. “Evacuees and Migrants Exhibit Different Migration Systems After the Great East Japan Earthquake and Tsunami.” *Demography* 57(4):1437–57. doi:10.1007/s13524-020-00883-7.

John Czajka and Amy Beyler. 2016. “Declining Response Rates in Federal Surveys: Trends and Implications.”

Johnson, Roger, Justin Bland, and Charles Coleman. 2008. “Impacts of the 2005 Gulf Coast Hurricanes on Domestic Migration The U.S. Census Bureau’s Response.”

Khaw, Khai Wah, Ramayah Thurasamy, Hadi Al-Abrrow, Alhamzah Alnoor, Victor Tiberius, Hasan Oudah Abdullah, and Sammar Abbas. 2021. “Influence of Generational Status on Immigrants’ Entrepreneurial Intentions to Start New Ventures: A Framework Based on Structural Equation Modeling and Multicriteria Decision-Making.” *Journal of Entrepreneurship in Emerging Economies* 15(3):589–634. doi:10.1108/JEEE-04-2021-0141.

Lathan, Hannah Stuart, Amy Kwan, Courtney Takats, Joshua P. Tanner, Rachel Wormer, Diana Romero, and Heidi E. Jones. 2023. “Ethical Considerations and Methodological Uses of Facebook Data in Public Health Research: A Systematic Review.” *Social Science & Medicine* 322:115807. doi:10.1016/j.socscimed.2023.115807.

Layne, Mary, Deborah Wagner, and Cynthia Rothhaas. 2014. “Estimating Record Linkage False Match Rate for the Person Identification Validation System.”

Lee, Everett S. 1966. “A Theory of Migration.” *Demography* 3(1):47–57. doi:10.2307/2060063.

Liebler, Carolyn A., Sonya R. Porter, Leticia E. Fernandez, James M. Noon, and Sharon R. Ennis. 2017. “America’s Churning Races: Race and Ethnicity Response Changes Between Census 2000 and the 2010 Census.” *Demography* 54(1):259–84. doi:10.1007/s13524-016-0544-0.

Linden, Thomas, Rishabh Khandelwal, Hamza Harkous, and Kassem Fawaz. 2019. “The Privacy Policy Landscape After the GDPR.”

Markovikj, Dejan, Sonja Gievska, Michal Kosinski, and David Stillwell. 2013. “Mining Facebook Data for Predictive Personality Modeling.” *Proceedings of the International AAAI Conference on Web and Social Media* 7(2):23–26. doi:10.1609/icwsm.v7i2.14466.

Massey, Douglas S. 1990. “Social Structure, Household Strategies, and the Cumulative Causation of Migration.” *Population Index* 56(1):3–26. doi:10.2307/3644186.

Massey, Douglas S. 2015. “A Missing Element in Migration Theories.” *Migration Letters* 12(3):279–99. doi:10.59670/ml.v12i3.280.

Massey, Douglas S., Joaquín Arango, Graeme Hugo, Ali Kouaouci, Adela Pellegrino, and J. Edward Taylor. 1993. “Theories of International Migration: A Review and Appraisal.” *Population and Development Review* 19(3):431–66. doi:10.2307/2938462.

Mayer, Jonathan R., and John C. Mitchell. 2012. “Third-Party Web Tracking: Policy and Technology.” Pp. 413–27 in *2012 IEEE Symposium on Security and Privacy*.

Mitchell, Travis. 2018. *Commercial Voter Files and the Study of U.S. Politics*. Pew Research Center. https://www.pewresearch.org/methods/2018/02/15/demographic-data/.

Molloy, Raven, Christopher L. Smith, and Abigail Wozniak. 2011. “Internal Migration in the United States.” *Journal of Economic Perspectives* 25(3):173–96. doi:10.1257/jep.25.3.173.

Mulrow, Edward, Ali Mushtaq, Santanu Pramanik, and Fontes Angela. 2023. “Assessment of the U.S. Census Bureau’s Person Identification Validation System.”

Mulry, Mary H., and Cristina J. Tello-Trillo. 2023. “Full Report of the Comparisons of Administrative Record Rosters to Census Self-Responses and NRFU Household Member Responses.”

Narayanan, Arvind, and Vitaly Shmatikov. 2007. “How To Break Anonymity of the Netflix Prize Dataset.”

Ortman, Jennifer, and Anthony Knapp. 2023. “Demographic Frame: Leveraging Person-Level Data to Enhance Census and Survey Taking.” Presented at the 2023 Southern Demographic Association Annual Meeting, San Antonio, Texas.

Piguet, Etienne, Pécoud Antoine, and Paul de Guchteneire. 2011. “Migration and Climate Change: An Overview.” *Refugee Survey Quarterly* 30(3):1–23. doi:10.1093/rsq/hdr006.

Piketty, Thomas, and Emmanuel Saez. 2014. “Inequiality in the Long Run.” 344(6186).

Porter, C. Christine. 2008. “De-Identified Data and Third Party Data Mining: The Risk of Re-Identification of Personal Information.” *Shidler Journal of Law, Commerce & Technology* 5:1. https://heinonline.org/HOL/Page?handle=hein.journals/washjolta5&id=40&div=&collection=.

Roth, Wendy D. 2009. “‘Latino before the World’: The Transnational Extension of Panethnicity.” *Ethnic and Racial Studies* 32(6):927–47. doi:10.1080/01419870802245042.

Stark, Oded, and J. Edward Taylor. 1991. “Migration Incentives, Migration Types: The Role of Relative Deprivation.” *The Economic Journal* 101(408):1163–78. doi:10.2307/2234433.

Stouffer, Samuel A. 1940. “Intervening Opportunities: A Theory Relating Mobility and Distance.” *American Sociological Review* 5(6):845–67. doi:10.2307/2084520.

Streiner, David L., Geoffrey R. Norman, and John Cairney. 2015. *Health Measurement Scales*. Oxford: Oxford University Press.

Taylor, J. Edward. 1984. *Differential Migration, Networks, Information and Risk*. Migration and Development Program, Harvard University.

United States Census Bureau. 2019. “Design and Methodology: Current Population Survey-- America’s Source for Labor Force Data Technical Paper 77.”

United States Census Bureau. 2021. “Group Quarters and Housing Unit Estimates Terms and Definitions.” https://www.census.gov/programs-surveys/popest/about/glossary/housing.html.

United States Census Bureau. 2025a. “2024 Survey of Income and Program Participation Users’ Guide.”

United States Census Bureau. 2025b. “About Metropolitan and Micropolitan Statistical Areas.” https://www.census.gov/programs-surveys/metro-micro/about.html.

United States Census Bureau. 2025c. “ACS Response Rates.” https://www.census.gov/programs-surveys/acs/.

United States Census Bureau. 2025d. “Decennial Census of Population and Housing Questionnaires & Instructions.” https://www.census.gov/programs-surveys/decennial-census/technical-documentation/questionnaires.html.

United States Census Bureau. 2025e. “Design and Methodology Report.” https://www.census.gov/programs-surveys/acs/methodology/design-and-methodology.html.

US Census Bureau. 2024. “2030 Census Research Recommendations: Inputs for the 2030 Census Operational Plan.” https://www.census.gov/newsroom/blogs/random-samplings/2024/12/2030-census-research-recommendations-operational-plan.html.

Wachter, Kenneth W. 2014. *Essential Demographic Methods:* Cambridge, MA: Harvard University Press.

Windzio, Michael. 2018. “The Network of Global Migration 1990–2013: Using ERGMs to Test Theories of Migration between Countries.” *Social Networks* 53:20–29. doi:10.1016/j.socnet.2017.08.006.

Xin, Rui, Niloofar Mireshghallah, Shuyue Stella Li, Michael Duan, Hyunwoo Kim, Yejin Choi, Yulia Tsvetkov, Sewoong Oh, and Pang Wei Koh. 2025. “A False Sense of Privacy: Evaluating Textual Data Sanitization Beyond Surface-Level Privacy Leakage.”

Zhou, Bing, Lei Zou, Ali Mostafavi, Binbin Lin, Mingzheng Yang, Nasir Gharaibeh, Heng Cai, Joynal Abedin, and Debayan Mandal. 2022. “VictimFinder: Harvesting Rescue Requests in Disaster Response from Social Media with BERT.” *Computers, Environment and Urban Systems* 95:101824. doi:10.1016/j.compenvurbsys.2022.101824.

Zou, Lei, Nina S. N. Lam, Heng Cai, and Yi Qiang. 2018. “Mining Twitter Data for Improved Understanding of Disaster Resilience.” *Annals of the American Association of Geographers* 108(5):1422–41. doi:10.1080/24694452.2017.1421897.

Zou, Lei, Nina S. N. Lam, Shayan Shams, Heng Cai, Michelle A. Meyer, Seungwon Yang, Kisung Lee, Seung-Jong Park, and Margaret A. Reams. 2019. “Social and Geographical Disparities in Twitter Use during Hurricane Harvey.” *International Journal of Digital Earth* 12(11):1300–1318. doi:10.1080/17538947.2018.1545878.