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: May 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).

One of the recent trends in data science is the demand for now-casting, or the ability to assemble data and generate data insights quickly. This is not possible with survey data, which can only describe the present after months of preparation and procedure. Then survey data rapidly loses its value for describing current conditions and needs another costly re-survey. An additional benefit of an administrative records approach is the reduction in measurement error. Only obtaining records removes some of the biases from self-report. Additionally, surveys can leverage administrative records by only having the respondent answer questions not available in the records.

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 correct 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). This method 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 primarily in obtaining stocks, not flows of people in particular geographies. An 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. The demographic frame utilizes a modeling strategy to match some 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, combining records across several years for a fixed 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 source to identify 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)*

*Operationalization of Place*

Researchers have tended to choose potential origin and destination locations based on the availability of data and the theoretical orientations held. 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 2025a). 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 smooth 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 area problem and, when ignored, leads to the ecological fallacy.

Federal data centers are a great source of data. Researchers can access to microdata, which is not aggregated at all and thus avoids the modifiable area problem. However, these researchers are limited by disclosure review board(s). Because of the anonymity concerns in identifying and tracking individual migrants across places, many analyses are aggregated up to a larger geography, and because of the modifiable area problem and the nesting of Census Bureau geographies into counties and state, these aggregations are usually counties or states.

There are theoretical reasons to consider the interactions of aggregates. 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 matrixes of migration from these aggregations (Curtis, Fussell, and DeWaard 2015; Hauer, Holloway, and Oda 2020; Johnson, Bland, and Coleman 2008). These matrixes often 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.

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 RESEARFHERS ^

*Operationalization of Time*

Researchers studying migration related to disasters have utilized a couple of different approaches to identify time periods for migration.

*Operationalization of Migrants*

*Data Sources Used for Disaster Research*

*Surveys.* The American Community Survey and the Decennial Census are like phone surveys in that all generate random samples from a sampling frame. The ACS and Decennial Census sample addresses to get their random samples. 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 correlations 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 American Community Survey 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 2025b), 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 American Community Survey in 2010 and in 2011 3.57 million addresses (households) are sampled each year to create the American Community Survey 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 disaster 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. 2010-2014, 2015-2019, etc.)

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).

*Administrative records.* 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 studies 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 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, age, sex, educational attainment, and more.

With access to IRS, ACS, and Census data, many of the drawbacks of a particular data source can be ameliorated. Using common identifiers for people and places, 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.

The first and biggest issue with using administrative records is matching respondents across different records (Harron et al. 2017). People changing their characteristics, like names, can make it difficult to match records collected for different purposes. Thankfully, the U.S. Census Bureau has a whole division working on the matching problem 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 was around .005% (Layne, Wagner, and Rothhaas 2014).

*Third-party records.* \*NEEDS FLESHING OUT\*

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. 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 have few levels tend to have higher accuracy than variables with more uniform spread and more levels.

If a person place assignment process could correctly include PIKs for 80% of the population and 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.

Having discussed PIKs and PIK variable assignment, the next consideration for person place assignment are the locations. The Census Bureau also has a solution for researchers here: the Master Address File has IDentification keys (MAFIDs) for addresses. The Master Address File is a record of all known addresses with people living in them, including group quarters, and is regularly updated. The American Community Survey avoids respondent-given addresses by using the Master Address File as their sampling frame. Administrative-data-based efforts rely on matching work to match respondent supplied addresses to Master Address File addresses. Fortunately, this work has been central to Geography division of the Census Bureau for many years. Many administrative data sources, like the United States Postal Service, also have their own MAFID matching processes.

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

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 Demographic Frame’s (demoframe) 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 1,000 sources including the U.S. Census Bureau’s version of the Social Security Administration’s Numerical Identification System (CNUM), data from the Bureau of Prisons, the U.S. Postal Service’s National Change of Address File, 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, 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 are 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 uses different 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 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 nearly a terabyte 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.

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 two 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. There are no other whole universe data products.

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. As a last step, I change retired MAFIDs to their ‘surviving MAFIDs’. Many of the multiple MAFID’s identified in the Person Place Table are MAFIDs that represented the same address at different points in time.

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. This is essentially an 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. I will calculate Cohen’s Kappa and Krippendorff’s Alpha, common measures of inter-rater reliability, between the various data frames.

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

*Analyzing the Migration Patterns of the 2020 Coronavirus Pandemic*

The 2019 demoframe extract provides a unique opportunity to observe the migration patterns during a particularly eventful time: the start of the 2020 Coronavirus Pandemic. Because the demoframe extract is a whole-universe accounting like the Decennial Census, we can see if and where people are moving by comparing the three datasets used in the previous analyses (the 2020v3 demoframe extract, the 2020 BRAPPA with retro- and pro-spective source utilization, and the 2020 Decennial Census Edited File) against the 2019v2 demoframe extract.

RESULTS

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| TABLE 1. Characteristics of 2020 Data Products. | | | | | |
| Data Product | Reference Date | Total Included PIKs | Total Included MAFIDs | PIK Coverage | MAFID Coverage |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |
|  |  |  |  |  |  |

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 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\_two\_source | April 1 2020 |  |  |  |  |
| BRAPPA\_four\_source | April 1 2020 |  |  |  |  |
| BRAPPA\_only\_sources | April 1 2020 |  |  |  |  |
| BRAPPA\_no\_vsgi | April 1 2020 |  |  |  |  |
| BRAPPA\_unitstat\_resstat | April 1 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 | Percent Agreement | Kippendorf’s Alpha | Cohen’s Kappa | Percent Agreement | Kippendorf’s Alpha | Cohen’s Kappa |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TABLE 4. Comparisons Between 2020 Data Products and the ACS | | | | | | |
|  | For Movers | | | For Non-Movers | | |
| Data Product | Percent Agreement | Kippendorf’s Alpha | Cohen’s Kappa | Percent Agreement | Kippendorf’s Alpha | Cohen’s Kappa |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TABLE 5. Comparisons Between 2020 Data Products and the BRAPPA | | | | | | |
|  | For Movers | | | For Non-Movers | | |
| Data Product | Percent Agreement | Kippendorf’s Alpha | Cohen’s Kappa | Percent Agreement | Kippendorf’s Alpha | Cohen’s Kappa |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| TABLE 6. Comparisons Between 2020 Data Products and the Random Forest PPM | | | | | | |
|  | For Movers | | | For Non-Movers | | |
| Data Product | Percent Agreement | Kippendorf’s Alpha | Cohen’s Kappa | Percent Agreement | Kippendorf’s Alpha | Cohen’s Kappa |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |

Chapter II: Estimating Migration

The purpose of this chapter is to connect the first and final chapters. We explore 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 business rules approach versus the more traditional person place model? I’ll again use the ACS and Decennial Census as benchmarks to argue for the applicability of the BRAPPA. 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.

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 at the same time.

REFERENCES

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.

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.

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*.

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.”

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.

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.

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.

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.

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.”

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.

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.

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.

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. 2025a. “About Metropolitan and Micropolitan Statistical Areas.” https://www.census.gov/programs-surveys/metro-micro/about.html.

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

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

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.