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 biases 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 will (1) assemble a dataset using unique identifiers maintained by the U.S. Census Bureau called Primary Identification Keys (PIKs). (2) I will rank the administrative records by how many correct PIKs they add, using the American Community Survey and the U.S. Decennial Census as separate comparators. (3a) 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. A researcher would normally include ACS or decennial data into the demographic frame as appropriate, but I will create a demo frame that avoids these rich data sources. This will offer a conservative test for the effectiveness of an administrative records approach. Afterwards, I will (3b) make comparisons with a demo frame that includes one of these sources compared to an omitted source. Having shown the ability to make estimates with the demographic frame, I will (4) estimate measures of migration, such as the in/out migration rates 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) combine the available datasets together to compare the migration related to hurricane Idalia.

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 and observe processes unseen by other methods.

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).

The next most important variable for a dataset examining migration 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.

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.

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. This project differs from the Demographic Frame in three ways. This work 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 (sometimes called business rules). The goal of the demographic frame is to provide a frame for the whole country that researchers can use easily. (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.

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). This project will supplement existing methods by describing the context of a time or place by using available administrative data. 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 Idalia 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.

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*. The World Bank.

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.

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

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.

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

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.

Layne, Mary, Deborah Wagner, and Cynthia Rothhaas. 2014. "Estimating Record Linkage False Match Rate for the Person Identification Validation System." Center for Administrative Records Research and Applications.

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

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

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.

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.

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.