Teagan Zuniga

October 13, 2023

COGS 212

Data Journey Narrative

The data that will be used for this project comes from Hauer & Byars (2019). Matthew Hauer is a sociologist at Florida State University whose work relates to migration and James Byars works in scientific computing in the Institute of Government at the University of Georgia. They created the dataset with the intention of it being used to study migration.

As the authors note, the IRS has released migration data since 1990.[1] The IRS estimates migration using tax filings. A person's tax filing in $year_i$ is compared to their tax filing in $year_{i+1}$. A person is considered to have migrated if their address in the $year_{i+1}$'s tax filing does not match the address $year_i$'s tax filing. What is released by the IRS is aggregate counts of how many people are considered to have migrated from $county_1$ to $county_2$ in a given year.[2] This allows scholars to explore county-to-county migration in the United States; however, there are three notable problems with this data.

First, this data is likely not fully representative of the U.S. population. Since the IRS is relying on tax filings, those who do not file taxes are underrepresented in the data (Gross, 2005; DeWaard, Curtis, & Fussell, 2016), particularly undocumented individuals, the poor, the elderly, and college students (Gross, 2005). However, given that the majority of U.S. householders file tax returns (Molloy, Smith, & Wozniak, 2011), this should not be a substantial problem, especially since I do not aim to answer questions related to *who* migrates.[3]

Second, while counties tend to be fairly stable geographic units, some do change over time. This data does not account for these changes. This is because the authors aimed to preserve the original data as much as possible – illustrating the importance of mutability/immutability of the

[1] The historic time series of the data was broken in 2011, as the IRS was improving their data processing methods, so this data only goes up to 2010.

[2] Cases where 10 or less people moved between two counties are removed from the dataset. This is to protect the anonymity of individuals.

[3] This IRS data does not include demographic data so it would be impossible to answer questions about *who,* even if the data was representative of the U.S. population.

data to the authors. Therefore, it may be more plausible to focus on state-to-state migration than county-to-county migration since state borders have not changed between 1990 and 2010.

Lastly, while the IRS migration data has great potential for the migration research community, the IRS posts the data in a very hard to manage format! The data is posted across over 2000 data files. Luckily, the goal of Hauer & Byars (2019) is to process this data and provide it to researchers as a single, flat file. While this is very helpful, it is important to note that this is an extra step in the data's journey and opens up the possibility for more error in the data. Luckily, Hauer & Byars (2019) post their R code, allowing users to check for any such errors.

As such, in a simplified form, the data journey can be illustrated as:

To restate, I have three interrelated research questions: 1) how has internal migration within the United States changed over time? 2) when people do migrate internally, where do they migrate from and 3) where do they migrate to? Therefore, the data provided by Hauer & Byars (2019) on U.S. county-to-county migration is suitable to answer my research questions.[4]

[4] There are other datasets that are frequently used to study subnational migration in the United States (the Decennial Census and the American Community Survey); however, those data sources rely on self-reported migration, are not in as readily usable formats, and rely on less concise temporal ranges. For example, the Decennial Census asks, "Where did you live five years ago?" This would not allow me to measure year-to-year migration. Therefore, I have chosen to use the processed IRS migration data provided by Hauer & Byars (2019).