Leah Silverman

**Motivation and Problem Statement:**

The dataset we will be exploring over the course of this project tracks the rates of migration between states within the U.S. as well as a variety of economic variables over the 16 years the data pull from. We will be investigating the correlation between these economic variables and the migration rates between different U.S. states. In addition to being an interesting exploration of human geography, being able to model the impact of these economic markers on people's personal choices around where to live has a huge impact on legislation and policy decisions. Migration processes, especially in large and varying amounts, have a wide ranging economic and social impact. Because of this, data around what does and does not cause this migration can greatly affect what economic issues lawmakers do and do not prioritize.

**Related Work**

I was brought onto the research team a few weeks ago to assist with statistical analysis and data exploration. So far, I have some limited variable exploration, focusing mainly on 5 different economic variables: average housing price, state gdp, economic freedom rating, and unemployment rate, across New England, Florida, California and Texas. I have done some visualizations surrounding how these variables change over the 16 year span in each state, and found the correlation coefficients between each of these variables and the in and out migration rates in each of the 9 states. I have not however looked at any intrastate trends.

I am not an expert in either the economics or public policy fields and do not have a full grasp on the depth of academic inquiry into the relationship between these fields and US migration patterns, but there’s certainly a large amount of research put into migration both within the US as well as other international systems. The economics professor who led the collection of this data believes that it is one of the first large data sets to explore the impact of these economic variables on migration patterns across the United States specifically. While this is incredibly exciting for the impacts of the research, it also makes it more difficult to find examples of models on this type of data. While the US census collects data on migration between counties, this data is more commonly used by local policy makers to analyze migration within states or regions, rather than across the country as a whole.

There is however, a large amount of research on in-region migration from rural to urban communities, and the impact of economic changes on this urban agglomeration. Most of the models focus on spatial relationships and use models such as the New Economic Geography model (NED) to describe movements of people through physical space (Baudino, 2021). This is a very different type of modeling than what we will be using on this data set, but is certainly still useful in the analysis of how the economy affects these smaller scale migration patterns.

The type of exploration we will be looking to do is much more commonly found in research on the migration patterns between EU nations. While there are many regulatory differences between the flows of migrants between US states, and EU countries, much of this research is incredibly applicable to the migration patterns we’re looking for. One such article, written by Arpaia et al. (2018) uses logistic regression models to track the effect of the business cycle on migration flows of workers between EU member nations. While a large part of their analysis focuses on the effect of this data on economic regulations within the European Union, their model is similar to the model we will likely use to explore similar patterns within the United States.

**Data Collection, Cleaning, and Exploration Plan**

To make things easier, all of our data has already been collected and pre-cleaned. It was originally uploaded in excel and stata, but I’ve read it into Jupyter notebooks as a csv. There are still a number of columns with NaN values that will need to be addressed, but that’s the only step left on the data cleaning side. As far as exploration goes, some very brief exploration has been completed, but a large chunk of the project will involve further analyzing the way the economic variables change over time, as well as the correlation between those variables and the rates of in-migration and out-migration. The data is organized by state pairs: each state has 49 entries per year, one for each of the other states.

**Modeling, Analysis, and Visualization Plan**

We hope to discover which variables affect migration inflows and outflows. In order to do this we will make a number of linear models modeling the relationships between different economic variables and the in and out migration rate for a variety of different states. Over the course of creating these models, we will hopefully find some correlation between changes in the economic variables and changes in the migration patterns that will allow us to predict how migration rates will change as these variables change. Our stretch goal will be to attempt to predict migration flows between certain states based on the economic variables given.

**Needs Assessment and Contingency Plans** a needs assessment and contingency plan, should identify shortcoming and gaps in your skills or knowledge that could be filled by additional class material, independent learning outside of class, or finding project teammates with complementary skills. This plan should also be explicit about the risks of the project and what could go wrong, and how you might address those issues

While I feel confident in my subject knowledge as it relates to the economic variables explored in the data set, I have very little previous experience with modeling or machine learning. I’m hopeful that as we continue to dive deeper into creating models and predictions this will quell some of these knowledge gaps, finding project partners who feel more confident in their ability to make and troubleshoot models could be very useful.

For the analysis, I do worry that we will struggle to find correlation between the economic variables we have and the rates of migration. Some of this may be mitigated by looking at multiple states at once and experimenting with groups of economic variables to attempt to find more clear correlations. Another potential issue with the data is that many of these economic variables are categorical in nature rather than numeric. This isn’t necessarily a pitfall but certainly can make exploration and analysis trickier. Luckily there is no text analysis though so we should generally be able to find work-arounds that still allow us to analyze the data.

**Proposed Timeline**

Week 1: In the first week we will focus on finalizing data exploration. We will discuss how to make a plan for how to deal with our remaining NaN values and find correlation coefficients for all relevant economic variables across all 50 states.

Week 2: In week two we will use the correlation coefficients we found to begin to make linear regression models, for select states and economic variables

Week 3: In week three we will finish the creation of these linear models and begin the troubleshooting process for these models. (back to step 1 of the data science life-cycle. (if we have time, we can also look to explore other variables and states, or some mix of both)

Week 4: We will use week 4 to continue to troubleshoot and analyze the models we began to create in week 2.

Week 5: If all goes smoothly timeline wise we will use the end of week 4 and week 5 to attempt to create models to predict the migration flow between states (our stretch goal). If not we will use these weeks to continue troubleshooting our models and analyzing both the issues these models have as well as the results they show

Week 6: finalize and write up findings and create presentation to present these findings to the class.

Works Cited

Arpaia, A., Kiss, A., Palvolgyi, B., & Turrini, A. (2018). The effects of european integration and the business cycle on migration flows: A gravity analysis. Review of World Economics, 154(4), 815-834. doi:https://doi.org/10.1007/s10290-018-0316-6

Baudino, M. (2021). Rural-to-urban migration in developing economies: Characterizing the role of the rural labor supply in the process of urban agglomeration and city growth. Annals of Regional Science, 66(3), 533-556. doi:https://doi.org/10.1007/s00168-020-01028-9

Pandit, K. (1997). Cohort and period effects in U.S. migration: How demographic and economic cycles influence the migration schedule. Association of American Geographers.Annals of the Association of American Geographers, 87(3), 439-450. Retrieved from https://login.ezproxy.uvm.edu/login?url=https://www.proquest.com/scholarly-journals/cohort-period-effects-u-s-migration-how/docview/208081480/se-2

U.S. Census Bureau. (2016) *County to-County Migration Flows* U.S. Department of Commerce. Retrieved from https://www.census.gov/data/tables/2020/demo/geographic-mobility/county-to-county-migration-2016-2020.html