ST441 Project Report:
Modeling Human Migration in the US

Anita Ruangrotsakun

I was inspired to do this project after reading Caleb Robinson and Bistra Dilkina's paper, "A Machine Learning Approach to Modeling Human Migration," in which they developed a machine learning framework to improve predictions for human migration patterns. I'm interested in learning more about machine learning and I would have liked to try to replicate their paper in full, but I knew it was a bit out of my reach in terms of knowledge and experience at this point in time. Knowing this, I decided to constrain the scope of my project to implementing only a gravity model and a radiation model in R in order to simulate migration patterns and compare them to historical data.

My original research objectives were as follows:

- Learn how gravity and radiation models are used to model human migration patterns.
- Learn how gravity and radiation models are implemented.
- Compare the results of gravity and radiation models to understand the different situations they are best suited for modeling.
- Learn how to use population data from the IRS and the US Census to discover and/or explore migration patterns.

Over the course of the term, I completed research on both gravity and radiation models for migration and different ways to implement them. I decided I would stick with the method outlined in Robinson and Dilkina's paper and Python library for migration as it would help me better understand their programming process and the steps I should take in my own code.

I then began sourcing data from the IRS and the US Census for the years 2009 through 2017. At this stage of the process, I learned that real-world data can be quite inconsistent and hard to find, though I eventually found suitable datasets online and wrangled them into a standardized tibble format.

After that, I set out to write a gravity model function that would produce a migration matrix as described by the mathematical explanation provided in Robinson and Dilkina's paper. I

quickly learned that I had to learn more about working with tibbles, matrices, and large number computations in R, but once I overcame that learning curve, I was pretty quickly able to write the functions I needed to implement both the gravity model and radiation model. You can view the resulting migration matrices produced by both of these models in the appropriate subdirectories of /results.

I then moved on to using these two models for comparison and simulation purposes. In the /analysis folder, you will find the simulate_future_migration file in which I found the most optimal alpha values to use for each model and ran both gravity and radiation models to migration flows for the years 2018 through 2022. I then created a function so that you could see the projected population growth for a given state over the years 2018-2022. In the /analysis folder, you will also find the compare_with_historical_data file in which I compute the residuals for the observed (historical) and predicted number of migrants for both models. I was at a loss for quite a while as to how I would visualize these results in an intuitive and easy-to-understand way, though I eventually settled on using a heatmap of residuals because it was pretty simple to implement from the resulting matrices I had and yet still interpretable.

Finally, I took actions to ensure my project would hit all elements outlined in the grading rubric. In the improve_efficiency file, you will find timing experiments and my attempts at better utilizing iteration and functions in order to run the gravity and radiation simulations multiple times using population data for different years. In /tests/testthat, you will find some unit tests I wrote to show that my functions work as intended and produce the expected results. Finally, in each of the files in /R, you will find all of the functions I wrote and their accompanying technical documentation.

I learned a lot about how to work with population data and modeling human migration over the course of this project. Should I continue working on this in the future, I would like to be able to implement the other two traditional models referenced in the paper (variations of the gravity and radiation model), as well as attempt to implement their machine learning model. I would also like to experiment more with different ways I could visualize the prediction and comparison results--perhaps using a geographic map of the US to visualize migration flows or projecting population growth for many states on a single graph--into a more comprehensive and potentially interactive visualization.