**Spatial and Longitudinal Data Analysis**

**Jose Zuniga**

**May 2018**

**Summary**

The IRS has a Spatial and Longitudinal dataset that contains U.S. Population Migration Data from 1990 to 2016. The data are “based on year-to-year address changes reported on individual income tax returns filed with the IRS.” These data contain geographic information regarding inflows and outflows to every county in the United States over time along with the number of returns, number of exemptions, and aggregate income associated with those flows. These data are very important to business for gauging demand in sectors such as marketing and real estate as well as supply of labor for human capital management.

These data are easy to find, but not easy to obtain or clean. The dataset is huge with inconsistent file formats, file naming conversions, file structures, and data values. On top of that, to pull one file the data store for the entire year must be downloaded. Data acquisition and cleansing was not trivial. Files were downloaded with [Bash](https://www.gnu.org/software/bash/), sifted manually, renamed with [Bulk Rename Utility](http://www.bulkrenameutility.co.uk/Main_Intro.php), reshaped with [Excel](https://products.office.com/en-us/excel), and then cleaned with [Python](https://www.python.org/) before being saved in an aggregated Spatial and Longitudinal CSV file that was uploaded to [GitHub](https://github.com/jzuniga123/SPS/blob/master/DATA%20608/IRS_NYc1990to2016io.csv).

With these data, an exploratory Shiny application was created that shows how people in the United States move in general. Yet the size of the data (millions of rows) makes the visualization unresponsive, therefore a subset of the data with inflows and outflows from New York (about a quarter-million rows) is used. The resulting visualization shows some interesting features. In general, New York residents tend to move between neighboring counties, but New York City residents also show a large amount of movement between other cities. Albany residents also have a wider dispersion than other New York residents, although not as wide as New York City residents. Without diving into a full quantitative analysis, at first glance this dataset appears to show that people tend to move between areas that are relatively similar in density. The interactive visualization produced through this Shiny application incorporates the following data visualization concepts:

1. **Data-Ink Ratio**: Simple color scheme, detail relative to zoom level, dynamic legend, minimalistic panel that fades.
2. **Chartjunk**: Map without gridlines, unnecessary colors, or elevations.
3. **Small Multiples**: Minimap (toggled) displaying spatial movement.
4. **Multifunctional Elements**: Choropleth polygons displaying density.
5. **High-Resolution Graphics**: Choropleth polygons represent highly dense data.
6. **Animation**: Year slider with play button to loops through time.
7. **Tooltips**: Display of Returns, Income, and Exemptions on hover.
8. **Zooming and Panning**: Can zoom with scroll wheel and pan by dragging.
9. **Web Controls**: Dropdowns, Radio Buttons, and Sliders.