Analysis of the Geographical Distribution of Socioeconomic Variables in the United States

Advised by: Dr. Hatfield

Senior Seminar

Department of Mathematics and Statistics

South Dakota State University

Braden Taubenheim

Spring 2020

Abstract

This research includes a spatial analysis of seven domains: race, sex, age, marital status, education, employment and insurance coverage. This analysis uses an individualistic approach of socioeconomic variables, as well as Moran's I to determine spatial autocorrelation.

Geographic Analysis of Socioeconomic Status in the U.S.

1 Introduction

Socioeconomic status can be defined as an individual's position in the social structure of today's society based on factors such as wealth, education, and income. Socioeconomic status is incredibly important in today's world, influencing future success of individuals as well as their children in many categories. Individuals with higher socioeconomic status have better access to health care, education, and safe housing. Socioeconomic status influences every aspect of life from the food one eats and the products they buy to their life expectancy and mortality rates. Socioeconomic variables can also be analyzed in a manner that shows social class trends in different regions of the United States. Median household income, education level, age, poverty rates, and health insurance status among other factors can show inequalities in certain areas and help provide a clearer view on the socioeconomic standing for a state. Understanding where states rank in various aspects of socioeconomic status can be beneficial for a number of reasons. It can help the government determine funding and the areas in which states are lacking most, help businesses make decisions on where to market certain products, and it can even be beneficial to individuals making life choices such as where to buy a new home or raise a child.

This paper will include a variety of socioeconomic variables across multiple analyses. This includes variables in the categories of race, sex, age, marital status, education, poverty rates and insurance coverage. This data covers the 48 contiguous states of the United States, gathered directly from the U.S. census website. Note that the following figures and calculations are computed in R.

2 Analysis by Socioeconomic Category

We will begin by analyzing critical individual socioeconomic categories. This will give insight into the importance of the variable and the distribution of the data. Some categories are divided into buckets, or different subdivisions of the variable. For example, the category education is subdivided into individuals with less than a high school diploma, a high school degree or equivalent, some college or an associate's degree, a bachelor's degree, or a graduate or professional degree.

Other categories are based on a single value, such as median household income within each state.

To begin the analysis of median household income, view the map below which indicates median household incomes for each state.

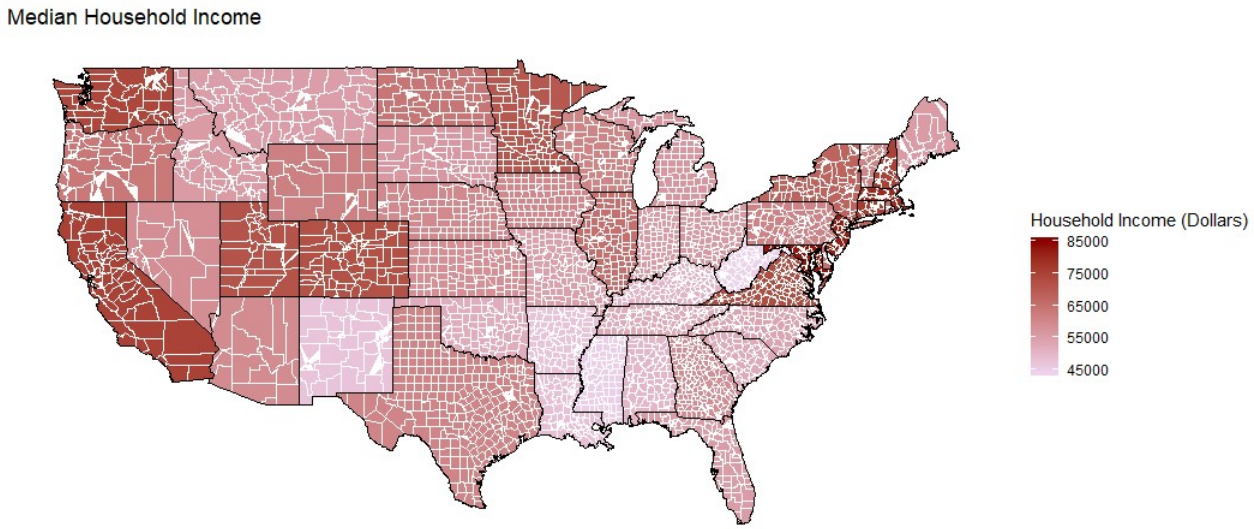

Figure 1: Median Household Income Map

It is interesting to note that the region with the highest median income is the District of Columbia at $85,203, while the lowest median income is in West Virginia at $44,097, and the average for the U.S. is $61,390. Investigating various state's median incomes could provide insight on the cost of living expenses, which may be helpful for families selecting their next state of residence. The standard deviation, which measures the amount of variation or dispersion of the variable, is $10,336.22. Another way to observe the dispersion of household income throughout the U.S. is to view the histogram and the box plot. Note that the following histogram shows a bimodal distribution, due to the high frequency of median incomes in the $55,000 and the $70,000 regions. Also observe that there are no outliers according to the box plot below.

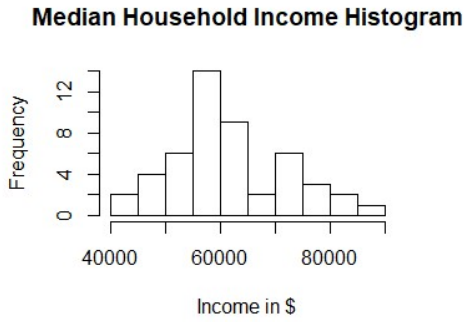

Figure 2: Median Household Income Histogram

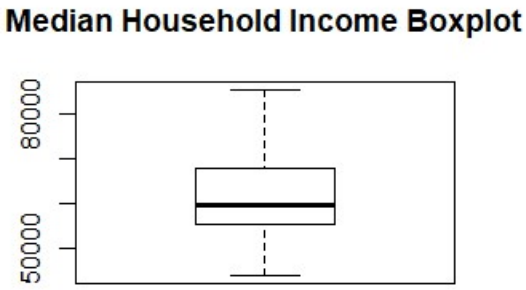

Figure 3: Median Household Income Histogram

The poverty rate for a state is a considerable measure of socioeconomic status as well. The U.S. Census Bureau uses a set of money income thresholds that depend on a household's size and composition to determine who is in poverty. Furthermore, if a household's total income falls below the determined income threshold, then that household is considered to be living in poverty. The determination of income thresholds is constant throughout the states, but it is updated frequently to account for inflation. The following figure is a map visualizing the poverty rates for the 48 contiguous states.

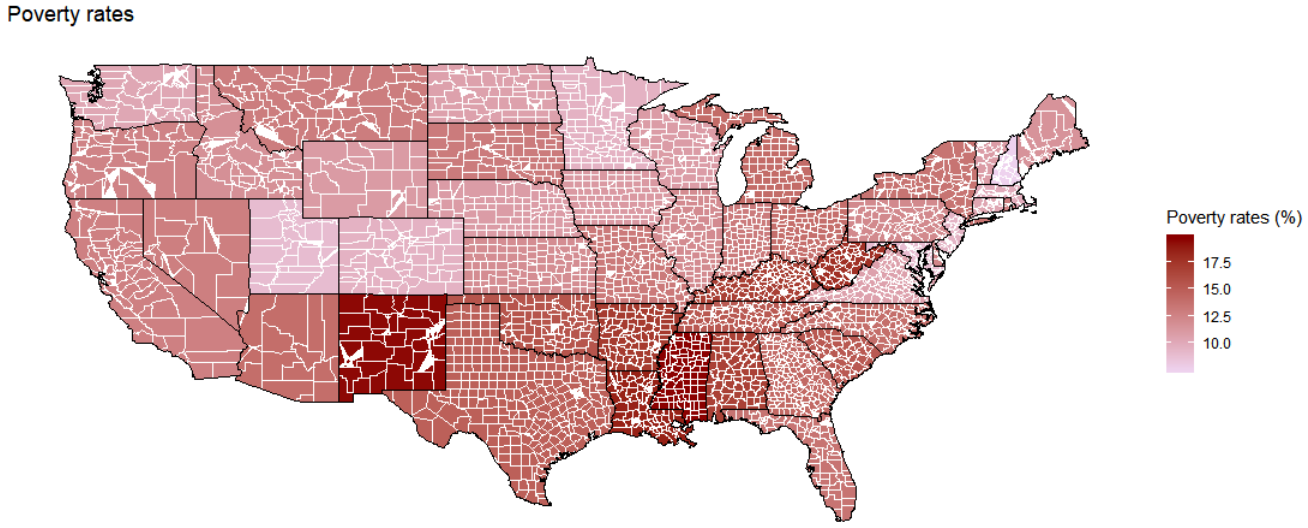

Figure 4: Poverty Rates Map

From figure 4, we obtain that Mississippi leads the U.S. in the percentage of households in poverty at 19.7%. New Hampshire brings in the lowest percentage at 7.6%, and the mean poverty

rate for the U.S. is 13.04%. A way in which these metrics could be beneficial, might be to assist federal government agencies in determining the allocations of their funds to fight poverty.

Another important socioeconomic variable is the median age for each state. The median age category will follow a similar analysis as that of the median household income category. Below is the map showing the median age for each state.

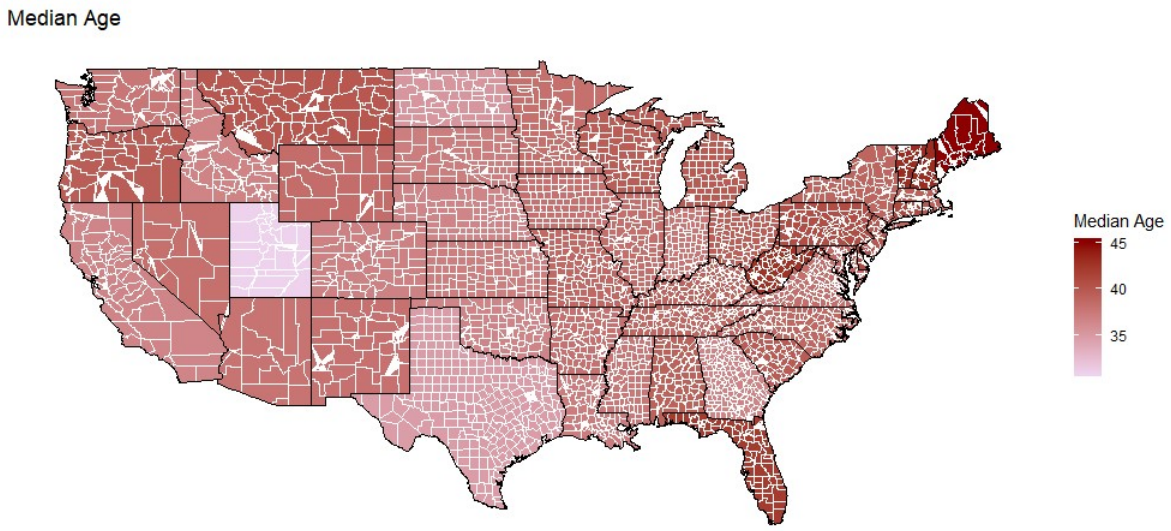

Figure 5: Median Age Map

Once again, noting a few metrics from the statistical summary; the maximum median age is 45.1 in Maine, the minimum median age is 31 in Utah, the average median age of the U.S. is 38.66 years, and the standard deviation is 2.4 years. This information provides insight on the demographic of a state, which could be helpful for various marketing practices. For example, a corporation looking to invest in senior living facilities may consider Maine as their best option due to the high median age of the state. The following histogram and box plot help visualize the dispersion of median age. The box plot exposes that the maximum and minimum median ages are outliers in this data set, while the histogram shows a relatively normal distribution of the data.

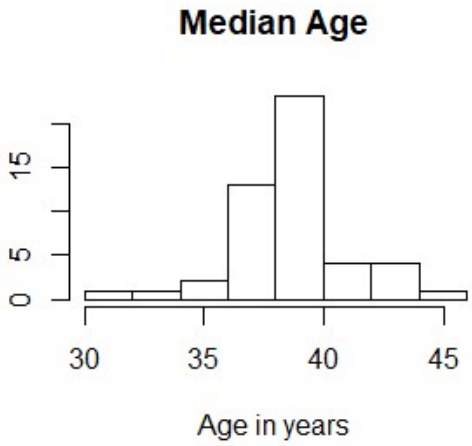

Figure 6: Median Age Histogram

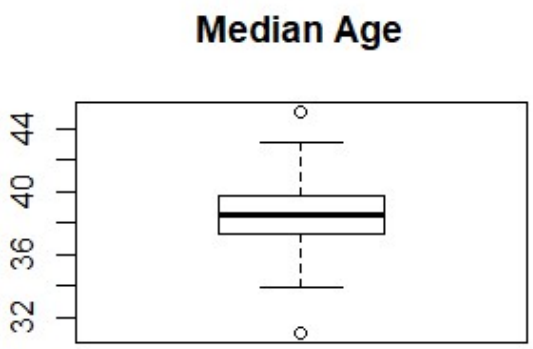

Figure 7: Median Age Box Plot

The next categories, education and health insurance coverage, are formatted in a bucket structure. Education has five different buckets: less than a high school diploma, high school graduate or equivalent, some college or an associate's degree, a bachelor's degree, and a graduate or professional degree. For a comparative view, we will analyze California, Minnesota, Nebraska, North Dakota, and South Dakota on a bar graph as a visual representation of the various education buckets for those states.

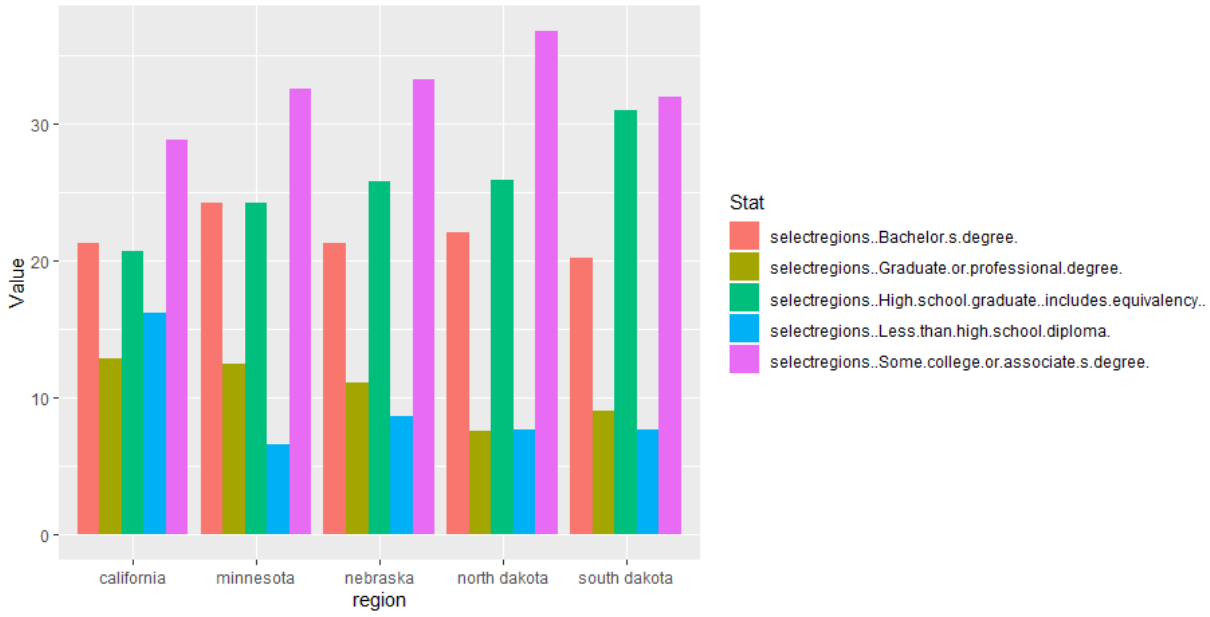

Figure 8: Education Bar Plot

Note that South Dakota has the highest percentage of high school graduates, yet they have the lowest percentage of individuals with a bachelor's degree. This could indicate an opportunity for increased university recruiting efforts. To further investigate education levels across the entire United States, observe the map below that indicates the percentage of individuals in a state that have a bachelor's degree or higher.

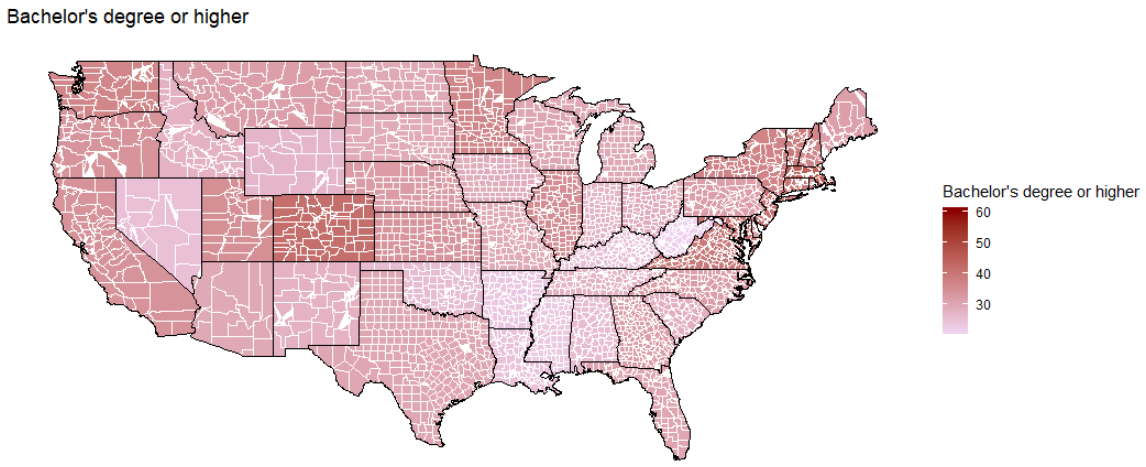

Figure 9: Bachelor's Degree or Higher Map

The state with the highest percentage of people with a bachelor's degree or higher is Massachusetts at 44.5%, the state with the lowest percentage is West Virginia at 21.3%, and the average percentage for our data set is 32.18%.

The last analysis of an individual category is for the health insurance coverage attribute. There are three buckets for health insurance coverage: private health insurance, public health insurance, and no health insurance. Once again, we compare these values for California, Minnesota, Nebraska, North Dakota, and South Dakota, visualized by a bar plot in figure 10.

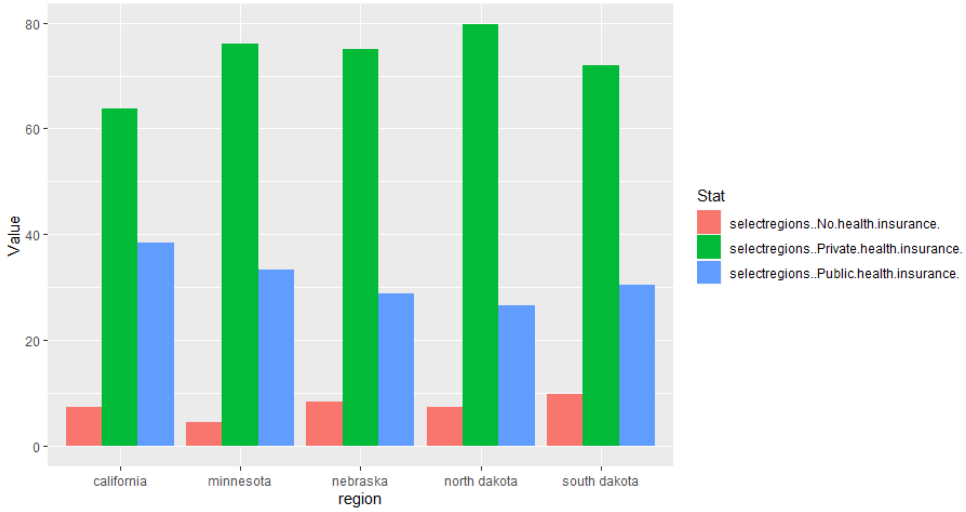

Figure 10: Health Insurance Bar Plot

Within the five selected states, it appears South Dakota has the highest percentage of individuals without health insurance at 10%. Furthermore, it is valuable to view the distribution of individuals without health insurance coverage across the United States to obtain a holistic perspective.

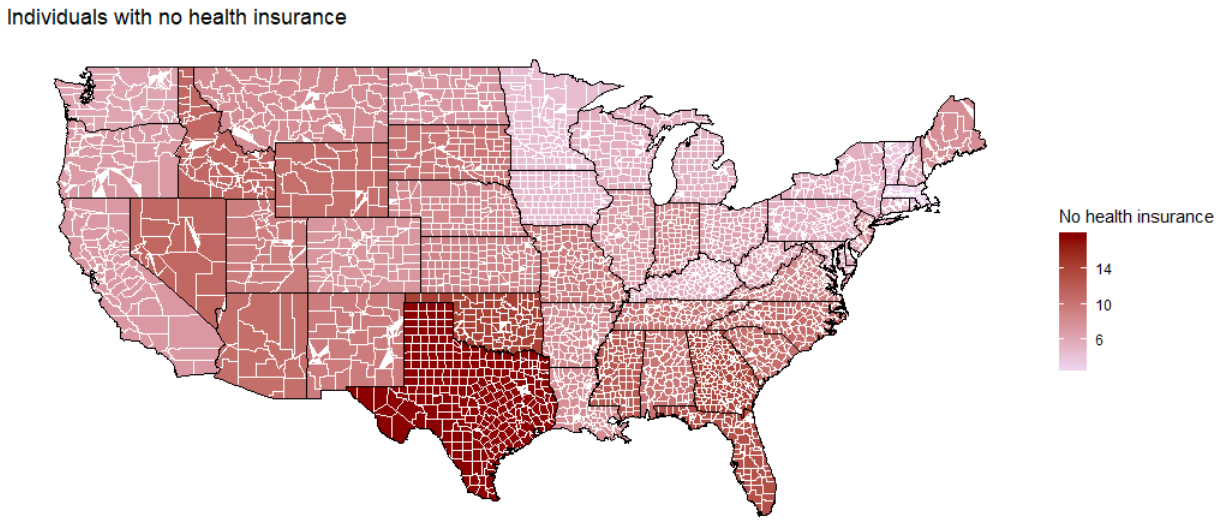

Figure 11: No Health Insurance Map

The maximum percentage of individuals without health insurance in a state is 17.7% in Texas, the minimum percentage of individuals without health insurance is 2.8% in Massachusetts, and

the average percentage is 8%. Analyses like these have proven to be helpful for government agencies regulating which states need the most medicaid assistance. This concludes our statistical breakdown of various individual categories. Next, we will view the spatial autocorrelation for select variables previously analyzed.

3 Spatial Autocorrelation using Moran's I

Spatial autocorrelation measures the degree as to which two observations, at different spatial locations, are similar to each other. Moran's I is the commonly used correlation coefficient to measure the overall spatial autocorrelation of the data set. This implies that Moran's I measures how one object is similar to others surrounding its location. Given a set of features and an associated spatial location, Moran's I evaluates the pattern expressed within the data set. In this context, Moran's I will be used to measure broad trends in our socioeconomic variables.

The calculation of Moran's I is determined by dividing the spatial covariation by the total variation. This can result in a vast range of positive and negative numbers. A positive Moran's I represents positive spatial autocorrelation, whereas a negative Moran's I represents a negative spatial autocorrelation, and a zero result represents no spatial autocorrelation.

Consider a region, R, subdivided into n cells, where each cell has a unique spatial feature. Then, the Moran's I statistic for spatial autocorrelation is given as:

$$I = \frac{\sum_i \sum_j w_{ij} cij}{s^2 \sum_i \sum_j w_{ij}} \tag{1}$$

Where $w_{ij} = 1$ if cells i and j are neighbors, $w_{ij} = 0$ if they are not neighbors. Also, note that $cij = (X_i - \bar{X})(X_j - \bar{X})$, where X_i and X_j are variables at two different locations, and the calculation for s^2 is as follows:

$$s^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2}{n} \tag{2}$$

Therefore, the Moran's I statistic is dependent on if the two objects are neighbors, the product of the difference between the respective values with the overall mean, and the sample size. The equation computes the mean and variance for the variable being evaluated. Then, for each spatial feature, it subtracts the mean, resulting in a deviation from the mean. Lastly, the deviation values for each neighboring spatial feature are multiplied then summed together. Once the Moran's I statistic is computed, we move on to evaluate the statistical significance of the test statistic. The

null hypothesis claims spatial randomness, indicating the spatial autocorrelation of the variable in question is zero. The statistical significance is based on a normal frequency distribution because most values cluster in the middle range of the data set and taper off in a symmetric fashion. Statistical significance for Moran's I is calculated as:

$$z = \frac{I - E(I)}{S_{error(I)}} \tag{3}$$

Where $E(I)$ is the expected Moran's I under the null hypothesis, and S is the standard error of the Moran's I value [2].

Using this method, we will compute the Moran's I values for each state and their corresponding p-value to determine the statistical significance. This will be conducted for three socioeconomic variables: median household income, median age, and the percentage of the those with no health insurance. Figures 12 and 13 help visualize the results for median household income.

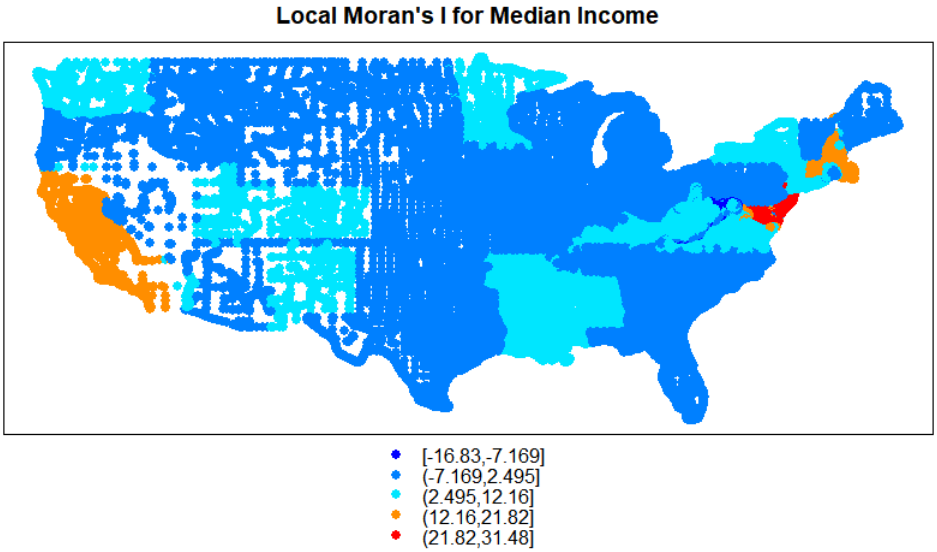

Figure 12: Local Moran's I values for median household income map

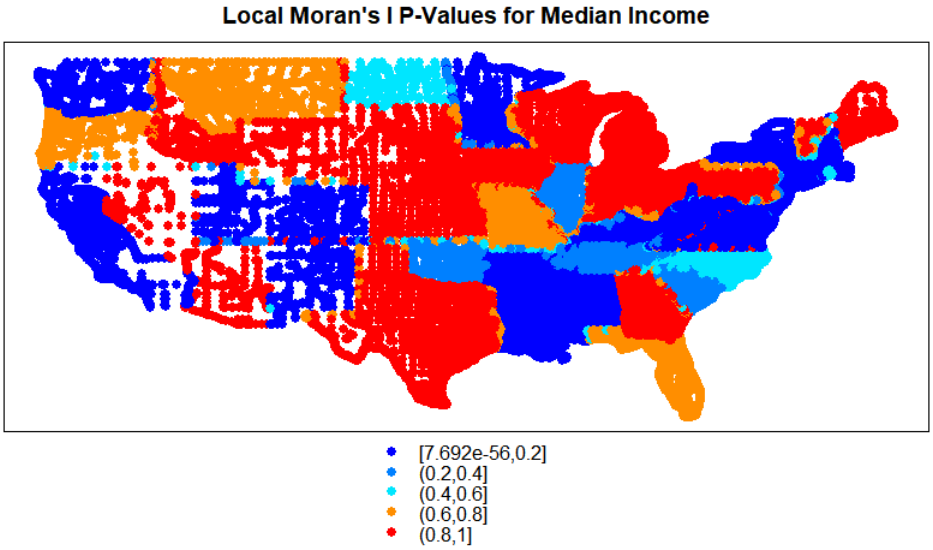

Figure 13: Local Moran's I P-values for median household income map

Observe that the majority of the U.S. shows a negative Moran's I value, with the exception of a few high Moran's I values in areas like the North-East region. Consequently, it follows that most regions have a high p-value as well. Using the recommended significance level of $\alpha = 0.2$ from the legend in figure 13, we reject the null hypothesis for areas in dark blue, and conclude that median household income is statistically significant and that there is strong spatial autocorrelation. In contrary to the dark blue regions, for areas in red, we fail to reject the null hypothesis at the significance level of $\alpha = 0.2$, and conclude median household income is statistically insignificant and spatially random.

Continuing on, figures 14 and 15 help visualize the results for median age.

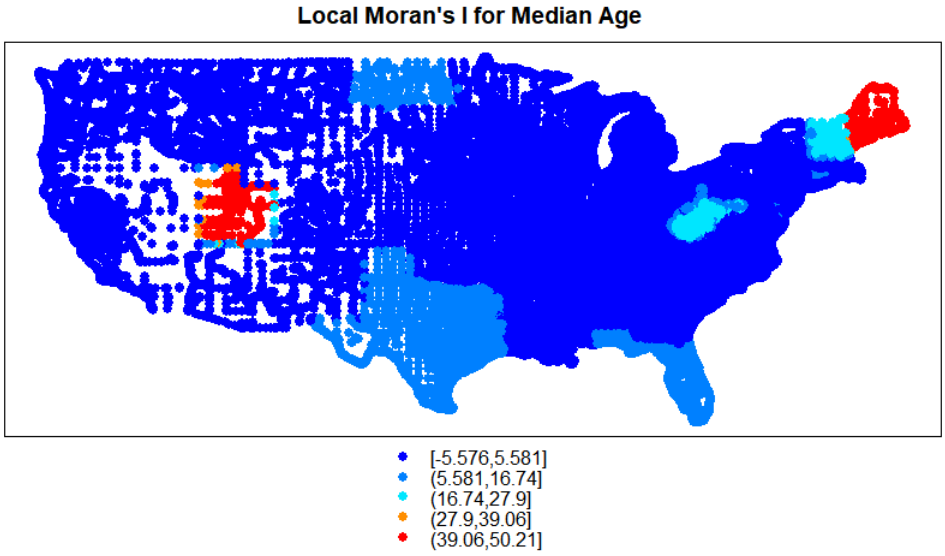

Figure 14: Local Moran's I values for median household income map

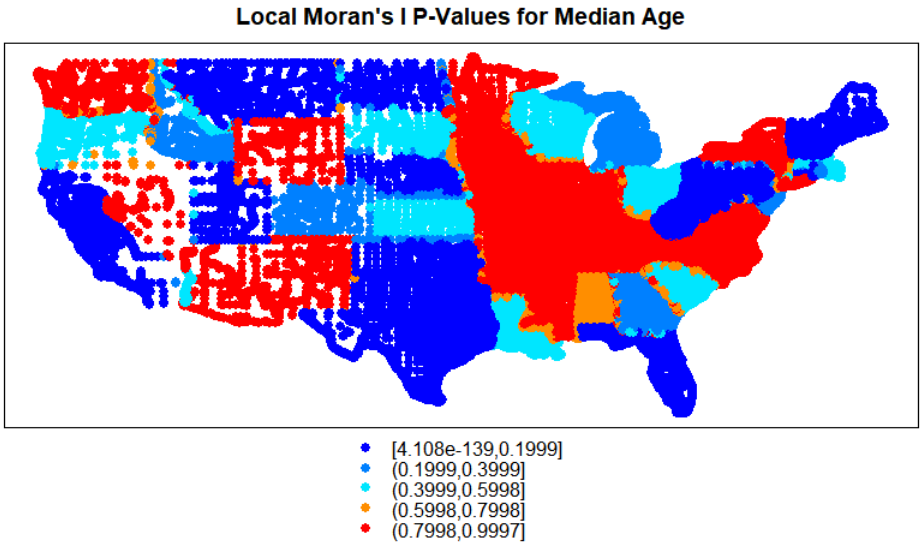

Figure 15: Local Moran's I P-values for median household income map

Once again, observe that even more of the U.S. shows a negative Moran's I value for Median age. Median age also results in less states with high p-values. Using a significance level of $\alpha = 0.2$, we reject the null hypothesis for regions in dark blue, and conclude that median age is statistically significant and there is strong spatial autocorrelation. For regions in red, we fail to reject the null hypothesis at the significance level of $\alpha = 0.2$, and conclude median age is statistically insignificant and spatially random.

Lastly, we will use the same visuals as before to determine the spatial autocorrelation of regions with high percentages of individuals without health insurance.

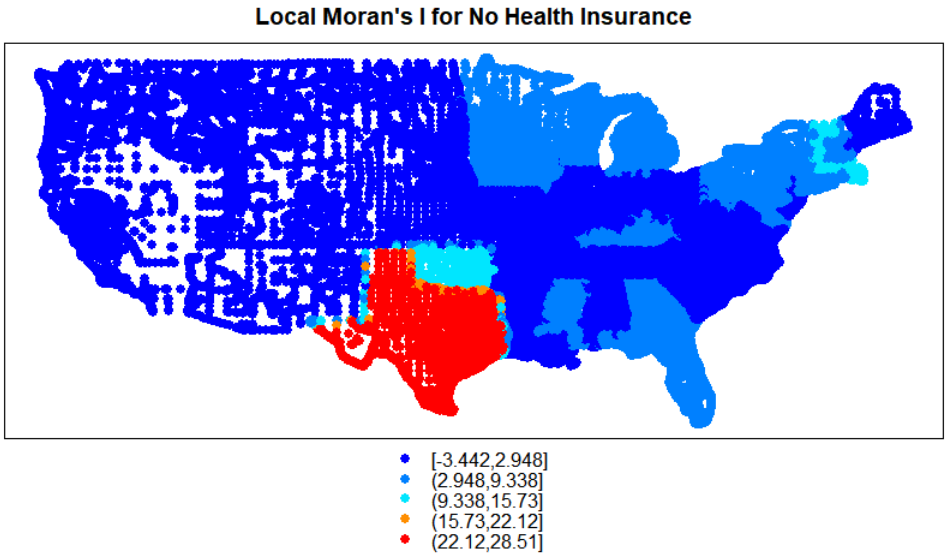

Figure 16: Local Moran's I values for percentages of no health insurance map

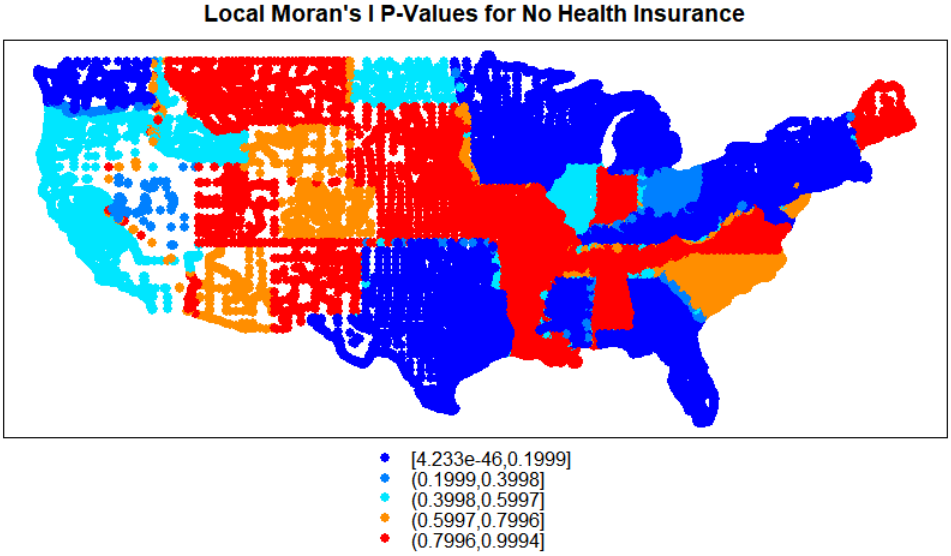

Figure 17: Local Moran's I P-values for percentages of no health insurance map

Figure 16 shows the majority of the contiguous U.S. has negative or low Moran's I values for the percentage of individuals without health insurance in that area, excluding the Texas/Oklahoma region. Figure 17 demonstrates the variety of p-values across the states. Once again, consider a

significance level of $\alpha = 0.2$. This results in the conclusion that there is statistical evidence to reject the null hypothesis in areas like the Minnesota/Wisconsin region, yet we fail to reject the null hypothesis in areas like the Nebraska/Kansas region. This implies there is spatial correlation in the former region and spatial randomness in the latter.

4 Conclusion

It is difficult to recognize which state has socioeconomic superiority given the complexity and the number of factors that contribute to socioeconomic standing. However, these findings can help shed light on a few key characteristics of socioeconomics among contiguous states. Socioeconomic status is comprised of many variables that critically affect lives in every regard. Analyzing a state's median income, poverty rate, median age, education level, and health insurance coverage can provide a comprehensive look on socioeconomics within that state. Another analysis could also infer which regions of the U.S. have a high spatial autocorrelation in the categories previously mentioned. Understanding these important aspects of socioeconomics within states and areas of the U.S. can be beneficial to a wide range of individuals, corporations, and administrations; from aiding government decisions on where to allocate state funding, to helping a couple choose where to start a family.

References

[1] Juan J. Abellan, Daniela Fecht, Nicky Best, Sylvia Richardson, David J. Briggs, "Bayesian analysis of the multivariate geographical distribution of the socioeconomic environment in England." Environmetrics 2007.

[2] Tonny J. Oyana and Florence M. Margai, *Spatial Analysis: Statistics, Visualization, and Computational Methods.* Florida: Taylor & Francis Group, LLC, 2016.

[3] R Core Team, "R:A language and environment for statistical computing." R Foundation for Statistical Computing, Vienna, Austria: 2013. http://www.R-project.org/

[4] United States Census Bureau, "Census Data." https://data.census.gov/cedsci/ (accessed December 3, 2019).

Braden Taubenheim is originally from Nebraska, but spent most of his childhood in Oklahoma and Minnesota. He is currently an undergraduate student at South Dakota State University and wrote this paper as part of the Senior Capstone requirements. Braden will complete his B.S. in mathematics in May 2020 and continue working at Wells Fargo as a Research Analyst. He hopes to rejoin the field of data analytics in the near future.

Braden knew he had a knack for mathematics when he continually beat the rest of his fourth grade class in the timed math competitions. He was always drawn to the rewarding nature of mathematics. The study of mathematics taught him a high level of problem solving that he will carry with him for the rest of his life.