2018

“What statistical and spatial relationships exist between health insurance, race, income, and education in the state of Georgia immediately before and after the implementation of the Affordable Care Act?”

EVAN A. WALKER

GEORGIA INSTITUTE OF TECHNOLOGY

Evan A. Walker

Capstone Project Paper

Summer 2018

Abstract

"What statistical and spatial relationships exist between health insurance, race, income, and education in the state of Georgia immediately before and after the implementation of the Affordable Care Act?". To answer this question, two datasets were used. They were both five-year estimates from the American Community sSurvey. The first range was for 2009-2013, and the second was an estimate from 2012-2016. The data obtained was for the 1959 census tracts in the state of Georgia. These years were chosen because the ACA was implemented in 2014, therefore the first dataset would not be affected by the ACA and the second would what largely be after its implementation. This study combined both linear statistical analysis as well as spatial statistical analysis. The variables chosen were income, race, education level, and health insurance. More specifically: average income for each tract, percent non-white/minority population, percent of individuals over 25 years-old with less than a high school diploma or GED equivalent, and the percentage of the population that in uninsured. These were chosen because I felt that they are all suitable metrics for examining these complex socio-economic factors. In the linear regression analysis health insurance was the dependent variable (DV) in all the regressions. For each dataset several combinations of the independent variables (IV) were used, in addition the difference between variables in the two time periods was regressed, and finally a logistic regression was performed on the differences between the two time periods. Unfortunately, the regression produced very little correlation amongst any of the variables.

(This will be discussed more thoroughly in the results section). The next part of the analysis was the spatial analysis for each variable a get-is Ord hotspot analysis was performed, a Moran's I test for spatial autocorrelation, and then individual choropleths were generated for each variable as well. In contrast to the linear regression there was significant correlation between the areas with high and low levels of each variable. This analysis was successful at examining the relationships of these variables spatially, statistically, and across time periods as well. Further research would be helpful using different benchmarks for race, income, and education that might produce more significant statistical results.

Introduction

Theodore Roosevelt was the first U.S. president to attempt healthcare reform in the United States in 1912. He proposed a system of national health insurance based on the German system of socialized medicine. He believed that healthcare for all was a basic human right like that of a public education. Teddy Roosevelt went on to lose the 1912 presidential election and thus the first attempt at healthcare reform was dead (Barr 2016). Ninety-eight years later The Affordable Care Act was passed on March 23, 2010 and was the biggest health care reform in the United States since the creation of Medicare and Medicaid in 1965 (Berlander 2010). In January of 2014 it was put into effect. Since then, the redaction of the ACA has been thrown around and used as a major talking point for gaining political capital. At the beginning of 2018 a huge push was made for the passage of the "American Healthcare Act" by the Trump administration, and despite an overwhelming Republican majority in both the Senate and House of Representatives, three republican senators refused to pass the legislation and no deal was made symbolizing a major defeat for the Trump administration. Since then every effort has

been made to dismantle the ACA and limit American's access to healthcare and deface Obama's signature legislation as much as possible.

According to the National Healthcare Expenditure annual report for 2016, it is estimated that $3.3 Trillion was spent on healthcare in the United States, coming out to $10,438 per person. This is an enormous cost for all Americans. Healthcare is not accessible to all Americans and health insurance is not available either.

Literature Review

As mentioned previously healthcare, health insurance, and their relationship with socioeconomic indicators is a very important topic that has been researched in the past. I reviewed three pieces of literature that fall under this "umbrella" in similar but different ways. The first two publications are similar in that they examine the role of SE indicators in access to healthcare, and furthermore how health insurance fits into that as well. The first paper focuses on the relationship between health insurance and healthcare access, and what role SE factors play in that relationship, while the second examines disparities in health behaviors amongst racial groups and how this translates to changes in mortality rate. The third study that was reviewed focuses on changes to the healthcare system because of the implementation of The Affordable Care Act or Obamacare. This group aimed to estimate the effect of three of the biggest facets of the ACA. They are premium subsidies for private coverage, the expansion of Medicaid, and the individual mandate, to quantify scientifically how much of the changes in health insurance pre and post ACA implementation can be explained by those factors. They used data from the American Community Survey for there analysis, which is the same data

source I used for my research. I felt that these three publications did a good job of representing current studies that are like the research I did for this project. They all examine some of the relationships and important aspects of healthcare and health insurance in the United States.

The first study I reviewed was titled “Health Insurance and Access to Healthcare in the United States” by Catherine Hoffman and Julia Paradise. This study examines the relationship between health insurance and access to health care. Throughout this report it is reiterated that health insurance, poverty and health are all interconnected. It states that 36% of America’s uninsured are people living in poverty. (Hoffman & Paradise) The study presents many profound statistics on health insurance and how America’s pluralistic health insurance system meaning it is both private and public with a basis being employer provided health coverage. (Hoffman & Paradise). The three primary public forms of health coverage in the United States are Medicaid, Medicare, and State Child Health Insurance Program (SCHIP). This study focuses primarily on Medicaid and SCHIP because Medicare is for people over 65. Without SCHIP and Medicaid an additional 61 million people in the US would not have health insurance. (Hoffman & Paradise). The conclusion of this study is that the number of uninsured people in the US is growing, their profile has not changed much over time, and the primary reason for people being uninsured is that they cannot afford health coverage. (Hoffman & Paradise.) This study looks at health insurance and its relationship to overall health. The fact that these researchers chose poverty as one indicator of health insurance and health outcomes. Therefore, I think this shows that income was a useful independent variable for this project.

The second study that I examined was titled: “Explaining US racial/ethnic disparities in health declines and mortality in late middle age: The roles of socioeconomic status, health

behaviors, and health insurance" by Joseph J. Sodano, and David W. Baker. This study focuses on the economic disparities amongst races in America and how they translate to mortality rates amongst different racial groups. In their study they wanted to examine how health status, health insurance, socio-economic status (SES), and health behaviors vary amongst racial groups, how much does racial/ethnic disparities explain these different health outcomes, and if this relationship is consistent. (Sodano & Baker) They performed a cohort study of a nationally representative sample of Hispanic, black, and white people aged 51-60. The study period was from 1992-1998. They performed statistical analysis using all these variables for the six-year time period. The study found there was a complex relationship amongst the variables, but that blacks and Hispanics has worse health outcomes then whites over the six-year period. The two predominate mediating causes of the health disparities were due to baseline health status and socio-economic status. (Sodano & Baker). This study used health insurance as the independent variable and examined how that effected things like mortality rate and SES disparities, while the study I completed used health insurance as the dependent variable. This study also used a cohort of almost 9,000 people over a 6-year period. This report presents some interesting relationships and shows that SES, race, and health are not independent of each other.

The third study I reviewed was titled: "Premium Subsidies, The Mandate, and Medicaid Expansion: Coverage Effects of the Affordable Care Act". It was published by Molly Frean, Jonathan Gruber, and Benjamin D. Sommers from the Bureau of Economic Research. This study provides the first comprehensive assessment of these provisions' effects, using the 2012-2015 American Community Survey and a triple-difference estimation strategy that exploits variation by income, geography, and time (Frean et al). The study used data from the 2012-2015

American Community Survey for individuals in the US that were under 65 and did not include Massachusetts in the data because of previously enacted legislation that had many of the same features as the ACA. The four dependent variables used to quantify health insurance were: no health insurance, Medicaid, employer sponsored insurance (ESI), and non-group private insurance. 98% of the US population fit into one of these four categories.

The researchers used different measures to quantify the three aspects of their study: premium subsidies, Medicaid expansion, and individual mandate. To quantify Medicaid expansion the researchers separated eligible individuals intro three categories: eligibility prior to the expansion, eligibility during early phases of the expansion (2011-2013), and newly eligible because of the full rollout. To measure premium subsidies the researchers calculated the unsubsidized premium for each housing insurance unit (HIU), which is the sum of the individual premium for each member of the household. Then the net premium was calculated for the remainder of the population. They then used these two measures to calculate percent subsidy which is the net premium divided by the unsubsidized premium. This was used to measure the effects of the subsidized premiums because of Obamacare. The mandate or the tax penalty associated with not having health insurance. This measure was applied to roughly 64% of the sample size for one reason or another and approached on average $1500 per family at higher incomes. (Fream et al).

The empirical design of the study allowed for a difference-in-difference-in-difference model across PUMAS, income groups, and time. Using 2012-2013 data as a control. (Fream et al). The results of the study show that the net increase of insured Americans across the study period was 40% due to premium subsidies for exchange coverage, and 60% due to

the expansion of Medicaid. Another key finding of the study was the lack of a “crowd-out” effect on private and employer sponsored insurance. This meaning that the increase in exchange provided insurance policies, and the expansion of Medicaid did not “crowd-put” private insurers from the insurance market. (Fream et al). The study also revealed small and insignificant effects of the individual mandate. The results of this study show that these key pieces of the ACA all had net positive effects on the number of insured Americans and did not produce the negative responses that are often used as criticisms for the ACA. Unfortunately, the ACA is still under fire and several key pieces of its existence have been drawn back due to the ignorance of leaders in today’s political climate.

These three pieces of research are similar to what I have done for this project and highlight the interest and importance of the topic I have chosen. The first two studies differ from mine in that they examine health outcomes and healthcare access along with health insurance, while my study does not consider healthcare access or health outcomes. The third study is very comprehensive in its analysis of the ACA and how that has affected health insurance coverage in the US. My research does not dive into the intricacies of the ACA as much but does highlight the increase in insured Georgians after its implementation. In summary these papers all attempt to explicate the relationships between the ACA, SE indicators and health insurance in the US, which was one of the aims of my research. While these are all academic publications published in worth sources, which my research is not. I feel that it does provide further insight into these relationships and furthers the discussion on this very important topic.

Data

The American Community Survey is an ongoing survey that supplements the decennial census. The ACS represents the short and long-form versions of the US Census that include detailed questions about population and housing characteristics. This survey is nationwide and aims to provide accurate and current data for demographics and socio-economic criteria on an annual basis (ACS Information Guide n.d.). The households that receive the American Community Survey form are chosen based on location to provide the most accurate and representative extrapolations of the data. Currently the ACS produces estimates at a 5-year and 1-year terms. Previously there was also a 3-year estimate produced, however this was discontinued ending in the year 2013. For this study two five-year estimates were used. The first was from 2009-2013, and the second was from 2012-2016. There were chosen based on their ranges straddling 2014 which was the year the ACA was fully implemented. Since these are both 5-year estimates they provide data for the entire population. I chose the census tract level for my geographic unit of measurement because data was provided for all 1959 census tracts in the state of Georgia. I downloaded the data via socialexplorer.com with a subscription provided by the Georgia Institute of Technology. As previously stated the three independent variables (IV) used in this study were: race, income, and education, while the dependent variable (DV) being health insurance. Data from the ACS is downloaded in a comma-delimited format, and there are over 30 different measures that can be chosen for each census tract. Figures 1 & 2 show the information downloaded for the two five-year periods.

Figure 1.

Figure 2.

These variables were downloaded as raw number as well as percentages of the population of each tract. Except for per capita income, I used the percentage as the value for my analysis. The variable I used for race was the percentage minority population for each tract. I calculated this my taking the percent white population and subtracting this from 100 and thus resulting in the non-white population percentage of each tract. I chose the non-white population as my variable because I suspected that minorities are more disproportionately uninsured than whites and this would produce greater significance in the statistical analysis. The educational attainment for population 25 years and older was separated into several different percentages based on level of education. I chose the lowest level, that being high school education or less. I chose this because lower education levels are associated with lower socio-economic status and thus would have a greater relationship to uninsured individuals. I used the dollar value of per capita income for each census tract. The DV in my statistical

analysis was health insurance. Similarly, to the education data there were several different factors related to health insurance. For my research I chose the uninsured category and the percentage of the population of each tract that were completely uninsured. I felt this would best represent those of lower SE status and be a good indicator of health insurance overall. The American Community Survey is a very valuable resource for any social science related research and provides accurate and up-to-date information about the US population. There were many other factors that could have been considered for my analysis such as health outcomes and specific racial discrepancies, however I wanted to do a lot of analysis on just four basic variables and thus limited my data to the variables discussed. The first phase of my analysis was liear regression comparing the IV's to the DV. For the spatial analysis portion of this project the census tract data in a CSV format had to be joined to a census tract shapefile. This join was done based on the FIPS code for each tract in the CSV and the corresponding code in the census tract shapefile. The Georgia census tract shapefile came from the Atlanta Regional Commission's Open Data Portal. Once this was completed each census tract had the corresponding race, income, education, and health insurance attributes associated with it. The first step of the analysis was the linear regression.

Linear Regression

"R is a language and environment for statistical computing and graphics." (R-Project) R statistical software was used to examine these relationships via linear regression modeling. Linear regression is a statistical method that tries to "explain" the effect that certain independent variables have on a dependent variable. For this study the dependent variable in each regression was the percentage of the population that was uninsured in each Georgia

census tract. The independent variables were the values used to measure race, income, and education for each census tract. To identify correlation between the variables I performed 16 different regression in R using several combinations of the variables from the two time periods. First, I tested each IV independently against the DV for both 2009-2013 and 2012-2016 data, and then all the IV's combined vs. the DV. I then performed regressions for each period using all combinations of sum and products between each variable for a combined equation. Next, I calculated the difference between the earlier and later periods for each variable in each census tract to produce "change" values for each variable to show how these values increased or decreased between the two five-year time periods. Using these "change" values I performed another regression with the change in health insurance vs. the sum of the changes in the IV's. I then set up a binary regression where a decrease in the uninsured population percentage corresponded to a 1, and an increase in the uninsured population corresponded to a 0. I tested this binary health insurance variable against all the change values for the IV's as well. I performed a forward stepwise regression on all the variables in the two time periods as well as the change variables also. Finally, I used the scale function to scale and center each variable to produce a standardized value for each change IV all combinations of each scaled IV were regressed against the binary change values. For each normal linear regression, I used R's ggplot 2 package to create a scatter plot of the IV vs. the DV in each equation as well as the best fit line and r-squared value for each equation as well. Figure 3 outlines the breadth and equations used for the linear regression portion of my project's analysis. Figure 4 also explains what the shorter variable names in the regression outputs correspond to. For the stepwise, binomial, and scaled variable equations I included the outputs of the regression results.

2009-2013 & 2012-2016 Data Standard Linear Regression Equations

- Health Insurance = Income
- Health Insurance = Minority Percentage
- Health Insurance = Education level
- Health Insurance = Income + Minority Percentage + Education Level
- Health Insurance = Income + Minority Population + Education Level + (Income x Minority Percentage) + (Income x Education Level) + (Minority Population x Education Level) + (Income x Minority Population x Education Level)

2009-2013 & 2012-2016 Stepwise Regression

- Health Insurance = Income + Minority Population + Education Level

"Change" Equations

- Health Insurance Change = Income Change + Minority Population Change + Education Level Change
- Binary HI Change = Income Change + Minority Population Change + Education Level Change

"Change" Stepwise

- Health Insurance Change = Income Change + Minority Population Change + Education Level Change

Scaled Variables Binary "Change"

- Binary HI Change = Scaled Income + Scaled Minority Population + Scaled Education Level + (Scaled(Income x Minority Percentage)) + (Scaled(Income x Education Level)) + (Scaled(Minority Population x Education Level)) + (Scaled(Income x Minority Population x Education Level))

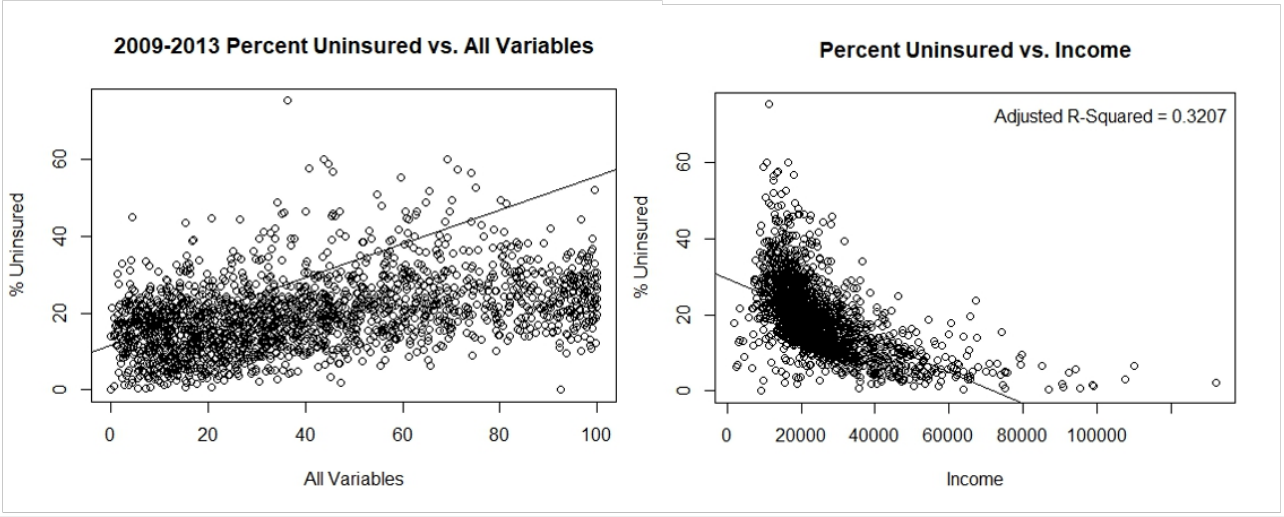
Figure 3

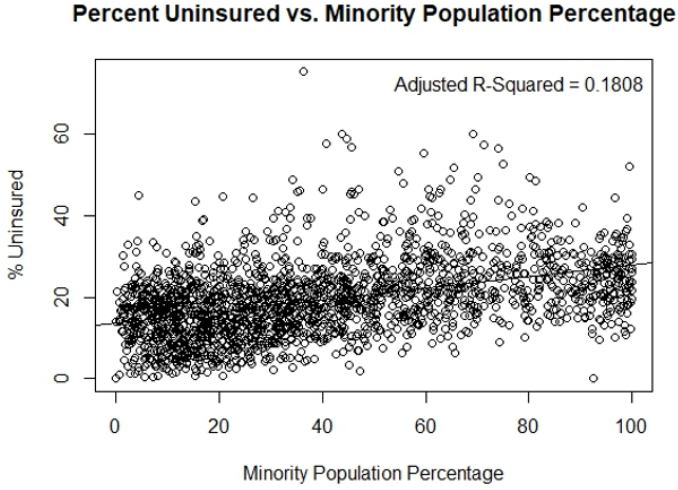
Regression output variable translations

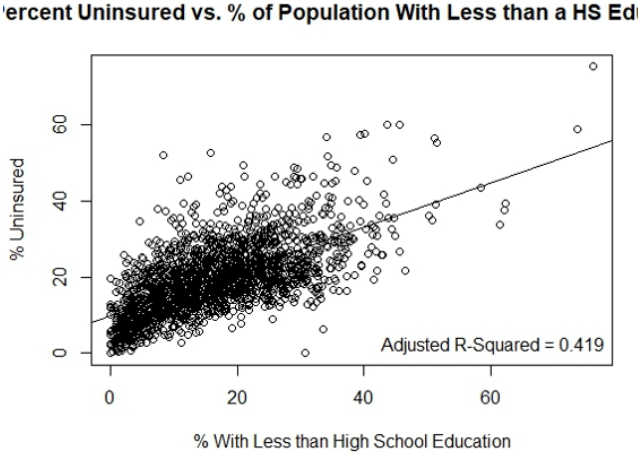
- **Less_HS** = 2009-2013 Percent of population with less than a high school education
- **HS_Less** = 2012-2016 Percent of population with less than a high school education
- **Income =** Average per capita income (same for 09-13 and 12-16)
- **PCT_MIN =** Minority population percentage (same for 09-13 and 12-16)
- **NO_HI** = percent of population with no health insurance
- **HS_Less_Ch =** difference between 12-16 and 09-13 percent of population with less than high school education
- **Income_Change =** change in per capita income between 09-13 and 12-16
- **PCT_MIN_Chng =** change in minority population between 09-13 and 12-16
- **PCT_HI_Change =** change in uninsured population between 09-13 and 12-16
- **Binary_HI =** change in uninsured population between 09-13 and 12-16 in which a value of 1 corresponds to a decrease in uninsured population and 0 corresponds to an increase.

Figure 4

2009-2013 Linear Regression

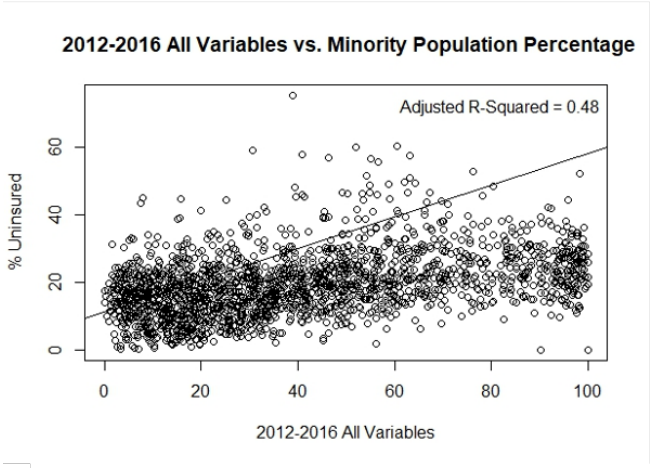


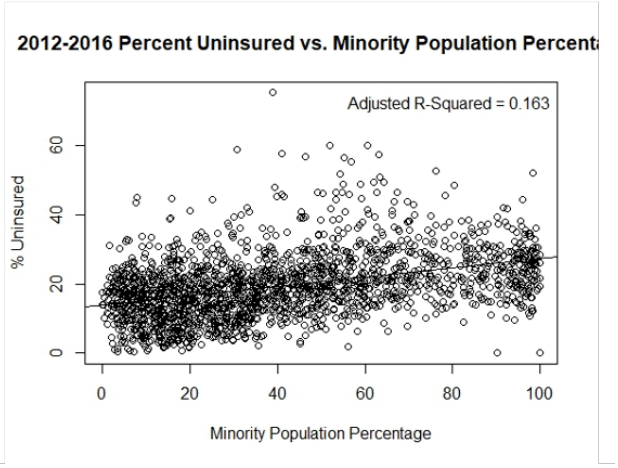


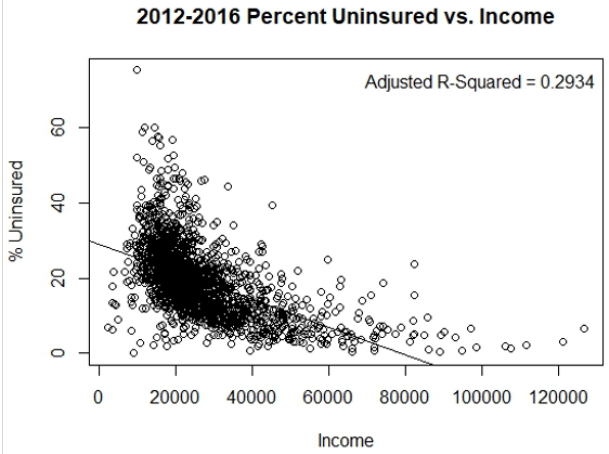


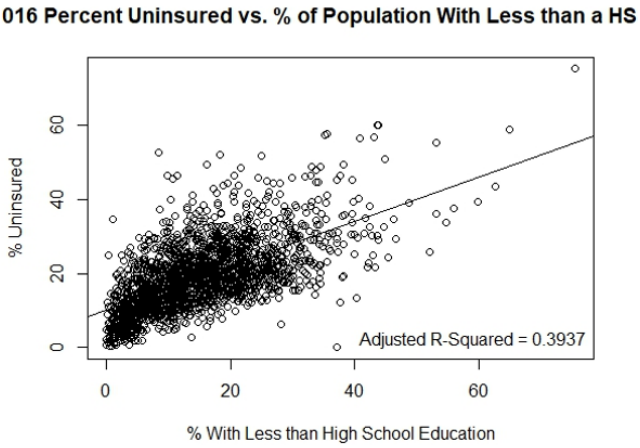

```
Coefficients:
                          Estimate Std. Error t value Pr(>|t|)
(Intercept)              9.162e+00  1.091e+00   8.397  < 2e-16 ***
Less_HS                  6.942e-01  6.008e-02  11.554  < 2e-16 ***
Income                  -6.392e-05  2.527e-05  -2.530  0.01150 *
PCT_MIN                  1.293e-01  2.630e-02   4.916 9.59e-07 ***
Less_HS:PCT_MIN         -7.213e-03  1.166e-03  -6.188 7.39e-10 ***
Less_HS:Income          -9.238e-06  3.176e-06  -2.909  0.00367 **
Income:PCT_MIN          -1.303e-06  9.510e-07  -1.370  0.17071
Less_HS:Income:PCT_MIN   3.617e-07  7.091e-08   5.100 3.72e-07 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.452 on 1946 degrees of freedom
  (15 observations deleted due to missingness)
Multiple R-squared:  0.5209,    Adjusted R-squared:  0.5192
F-statistic: 302.3 on 7 and 1946 DF,  p-value: < 2.2e-16
```

2012-2016 Linear Regression










```
Call:
lm(formula = NO_HI ~ HS_Less + Income + PCT_MIN + HS_Less * PCT_MIN +
    HS_Less * Income + Income * PCT_MIN + Income * PCT_MIN *
    HS_Less, data = myData)

Residuals:
    Min      1Q  Median      3Q     Max
-28.823  -4.220  -0.711   3.529  31.990

Coefficients:
                        Estimate Std. Error t value Pr(>|t|)
(Intercept)            7.635e+00  1.125e+00   6.788 1.50e-11 ***
HS_Less                6.399e-01  6.456e-02   9.912  < 2e-16 ***
Income                -4.808e-05  2.475e-05  -1.943  0.05217 .
PCT_MIN                1.620e-01  2.590e-02   6.256 4.83e-10 ***
HS_Less:PCT_MIN       -6.521e-03  1.249e-03  -5.221 1.97e-07 ***
HS_Less:Income        -1.284e-06  3.346e-06  -0.384  0.70123
Income:PCT_MIN        -1.504e-06  8.989e-07  -1.674  0.09438 .
HS_Less:Income:PCT_MIN 2.265e-07  6.922e-08   3.272  0.00109 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 6.601 on 1946 degrees of freedom
  (17 observations deleted due to missingness)
Multiple R-squared:  0.4984,    Adjusted R-squared:  0.4966
F-statistic: 276.2 on 7 and 1946 DF,  p-value: < 2.2e-16
```

2009-2013, 2012-2016, "Change" Stepwise Regression

```
NO_HI ~ Less_HS + Income + PCT_MIN

          Df Sum of Sq    RSS    AIC
<none>                  83910 7354.8
- Income   1    1794.9  85705 7394.1
- PCT_MIN  1    8289.2  92199 7536.9
- Less_HS  1   23430.6 107341 7834.0
> step$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
NO_HI ~ Less_HS + Income + PCT_MIN

Final Model:
NO_HI ~ Less_HS + Income + PCT_MIN


  Step Df Deviance Resid. Df Resid. Dev      AIC
1                       1950    83910.2 7354.783
```

```
Start:  AIC=7442.8
NO_HI ~ HS_Less + Income + PCT_MIN

          Df Sum of Sq    RSS    AIC
<none>                  87776 7442.8
- Income   1    1469.4  89246 7473.2
- PCT_MIN  1    8790.5  96567 7627.3
- HS_Less  1   24484.7 112261 7921.6
> step$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
NO_HI ~ HS_Less + Income + PCT_MIN

Final Model:
NO_HI ~ HS_Less + Income + PCT_MIN


  Step Df Deviance Resid. Df Resid. Dev      AIC
1                       1950   87776.26 7442.799
```

```
Start:  AIC=7185.23
PCT_HI_Change ~ HS_Less_Change + Income_Change + PCT_MIN_Chng

                Df Sum of Sq   RSS    AIC
<none>                       75387 7185.2
- HS_Less_Change  1    622.98 76010 7199.4
- PCT_MIN_Chng    1   1269.97 76657 7216.1
- Income_Change   1   2038.53 77426 7235.8
> step$anova
Stepwise Model Path
Analysis of Deviance Table

Initial Model:
PCT_HI_Change ~ HS_Less_Change + Income_Change + PCT_MIN_Chng

Final Model:
PCT_HI_Change ~ HS_Less_Change + Income_Change + PCT_MIN_Chng


  Step Df Deviance Resid. Df Resid. Dev      AIC
1                       1965    75387.5 7185.233
```

"Change": All variables, Binary all variables, scaled IV's all combinations

```
Call:
lm(formula = PCT_HI_Change ~ HS_Less_Change + Income_Change +
    PCT_MIN_Chng, data = myData)

Residuals:
     Min       1Q   Median       3Q      Max
-101.910   -2.917   -0.183    2.988   94.319

Coefficients:
                Estimate Std. Error t value Pr(>|t|)
(Intercept)     2.525e+00  1.552e-01  16.263  < 2e-16 ***
HS_Less_Change -1.349e-01  3.346e-02  -4.030 5.80e-05 ***
Income_Change   2.914e-04  3.997e-05   7.289 4.49e-13 ***
PCT_MIN_Chng   -1.349e-01  2.346e-02  -5.753 1.01e-08 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 6.194 on 1965 degrees of freedom
  (2 observations deleted due to missingness)
Multiple R-squared:  0.05856,   Adjusted R-squared:  0.05712
F-statistic: 40.74 on 3 and 1965 DF,  p-value: < 2.2e-16
```

```
Call:
glm(formula = Binary_HI ~ HS_Less_Ch + Income_Cha + PCT_MIN_Ch,
    family = "binomial", data = LogisticData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4644   0.3942   0.6922   0.7844   1.2278

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.14477    0.05387  21.252  < 2e-16 ***
HS_Less_Ch  -0.25194    0.05512  -4.571 4.86e-06 ***
Income_Cha   0.27872    0.05665   4.920 8.65e-07 ***
PCT_MIN_Ch  -0.04257    0.05223  -0.815    0.415
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2210.7  on 1970  degrees of freedom
Residual deviance: 2156.1  on 1967  degrees of freedom
AIC: 2164.1

Number of Fisher Scoring iterations: 4
```

```
Call:
glm(formula = Binary_HI ~ HS_Less_Ch + Income_Cha + PCT_MIN_Ch +
    HS_Less_Ch * Income_Cha + HS_Less_Ch * PCT_MIN_Ch + Income_Cha *
    PCT_MIN_Ch + HS_Less_Ch * PCT_MIN_Ch * Income_Cha, family = "binomial",
    data = LogisticData)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4139   0.3951   0.6859   0.7813   1.2427

Coefficients:
                                 Estimate Std. Error z value Pr(>|z|)
(Intercept)                       1.15186    0.05464  21.082  < 2e-16 ***
HS_Less_Ch                       -0.25715    0.05677  -4.530 5.90e-06 ***
Income_Cha                        0.26123    0.05912   4.419 9.94e-06 ***
PCT_MIN_Ch                       -0.06171    0.05458  -1.131    0.258
HS_Less_Ch:Income_Cha             0.06347    0.06131   1.035    0.301
HS_Less_Ch:PCT_MIN_Ch             0.07343    0.05804   1.265    0.206
Income_Cha:PCT_MIN_Ch             0.01892    0.05954   0.318    0.751
HS_Less_Ch:Income_Cha:PCT_MIN_Ch -0.07758    0.06308  -1.230    0.219
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2210.7  on 1970  degrees of freedom
Residual deviance: 2151.6  on 1963  degrees of freedom
AIC: 2167.6

Number of Fisher Scoring iterations: 4
```

Linear Regression Discussion

The two main indicators I looked at in determining the effectiveness of each regression model was the p-value associated with each variable and the adjusted R-squared value for each model as well. A general rule is statistical analysis is that a p-value of less than 0.05 or 5% means you can reject the null hypothesis and therefore that variable is significant is a meaningful addition to the model (How to Interpret Regression Analysis Results: P-values and Coefficients). R will conveniently denote symbolize these significant variables with an asterisk in the regression output, where more asterisks means a lower p-value starting with one asterisk for a less than 0.05 value, all the way to 3 asterisks for a p-value of 0. Each model that tested the IV vs. the DV individually showed p-values of zero, for both the 2009-2013 models as well as the 2012-2016 models. Additionally, in the models that included all the variables every IV had a p-value of zero indicating significance. In the 2009-2013 combined variable model all of the variable combinations, with the exception of Income x Minority Population, had p-values less than 0.05. In the 2012-2016 combined variables model all the variable combination showed significance except for Income, Less than high school education x Income, and Income x Minority population percentage. All the variables were significant in the change model. This was also true for the binary change model. The binary combined scaled variables model only showed significance with the IV's by themselves vs. their respective combinations. Based on these results the majority of IV's chosen for this analysis were significant and meaningful in their respective models.

However, the adjusted R-squared values and Akaike Information Criterion (AIC) numbers were not as significant. The graphs include scatterplots of the listed variables as well as the line

of best-fit and adjusted r-square value. The higher the adjusted r-squared value the more correlation there is between variables, and a steeper line of best-fit indicated strong correlation as well. This was not the case for any of the equations I used. The highest adjusted r-squared values achieved all came when it was a combination of the variables. Whether that be the sum of all the variables or the combination of the sums and products of all variables. All of them were close to 0.50 with the greatest being 0.5192, which was achieved using all sum and product variations in the IV's for the 2009-2013 period. The lowest out of this group were from the 2012-2016 equation using the sum of all the variables. This equation had an r-square value of 0.48. Adjusted R-squared could be interpreted as the percentage of explanation an IV has on the DV, so even the best adjusted r-squared value from the analysis of 0.5192, is barely better at explaining the DV than flipping a coin.

Testing the individual variables vs. the dependent variable produced even lower adjusted R-squared values. The highest out of this group was 0.419 which was the result of 2009-2013 education vs. health insurance. The lowest adjusted r-squared value was 0.163 which was 2012-2016 minority population percentage vs. uninsured. Finally, out of the standard linear regressions the sum of all the change values in each variable vs. the change value in health insurance resulted in an adjusted r-squared value of 0.057 which is very low.

To increase the breadth of my statistical analysis, I decided to perform a forward stepwise regression on the 2009-2013, 2012-2016, and change values to test which of the variables in the model had the most significant impact on the DV. When doing a stepwise regression an Akaike information criterion (AIC) is produced for each variable. The AIC determines which model or variable in this case is best for testing correlation. A lower AIC

means that the model is better. (AIC Wikipedia) None of the models or variables produced an AIC of less than 7100. The change values seemed to produce the best AIC with a score of 71185. The worst was the 2012-2016 data. These high AIC scores show that these models were not very good for testing this set of variables.

Next, I performed a binary logit regression where an increase in the uninsured population of a census tract was given a value of 0, and a decrease was given a value of 1. Binary logit regression is used to test these types of DV's in which there are only two outcomes. These outputs also produced AIC's for each equation. Using the DV in a binary format compared the sum of the changes in each of the IV's I was able to get an AIC of 2164, which is significantly better than the stepwise results, but still not great. Finally, to produce some measure of correlation I used a scale function on the change values in each IV. This centers and scales the variables. Centering the variables means that the arithmetic mean is subtracted from each value producing an average of "0". (SPSS Tutorial). Scaling the variables standardizes them by dividing the centered value by their standard deviations. (R-Documentation) A binary regression was then done using the binary health insurance change values against the sum of all scaled IV sum and product combinations. This produced an AIC of 2167, which again implies very little correlation between the variables.

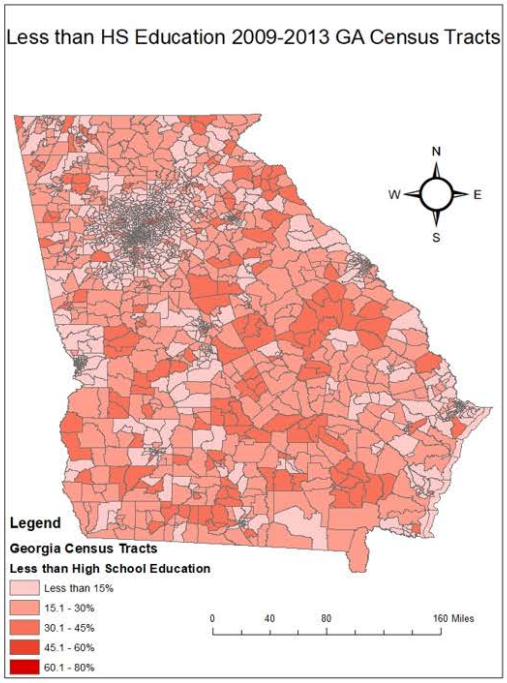
After rigorous statistical analysis all of the models showed one or more significant variables having p-values less 0.05. Unfortunately after interpreting the adjusted R-squared values and AIC these variables did not show much correlation at all on any level. I feel the combination of different regression types and equations thoroughly tested for correlation and

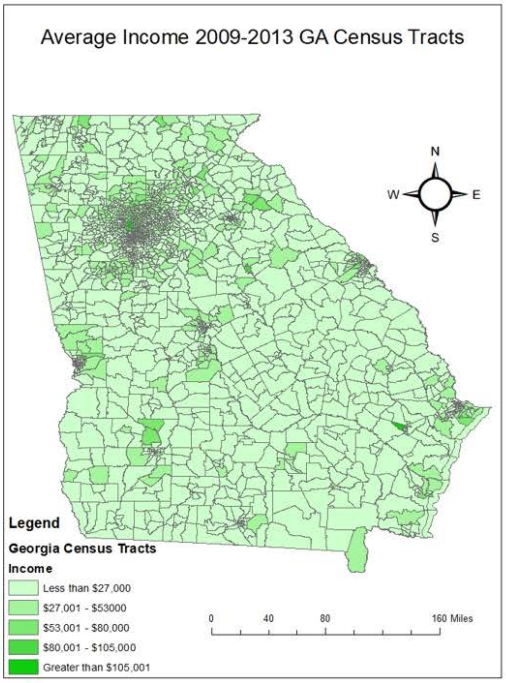
shows that these variables in the current form are not very correlated. If any conclusions can be drawn it may that the 2009-2013 data showed more correlation than the 2012-2016 data. Different measures of these variables should be used for further research, and the spatial relationships should be compared as well, which is what leads to the next section of this analysis.

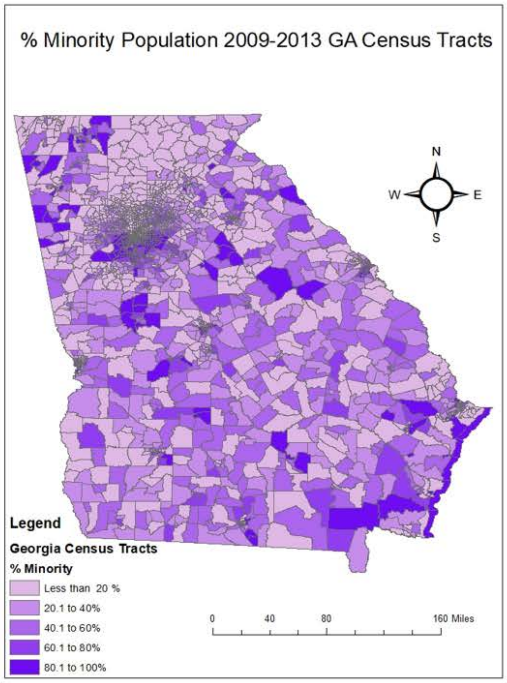
Spatial Analysis

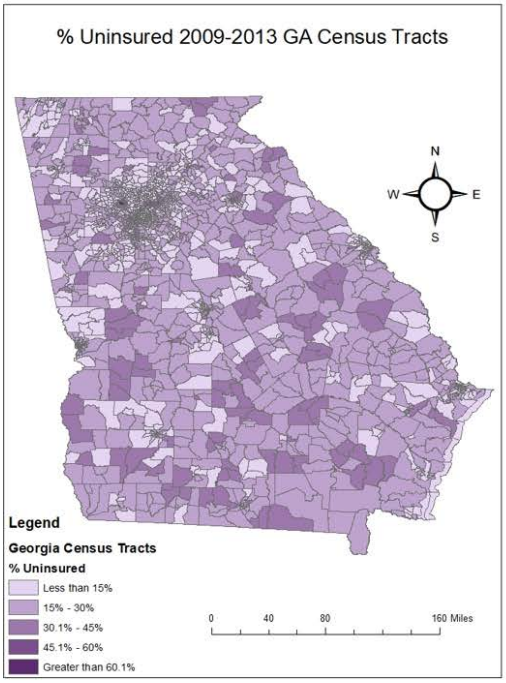
To analyze the spatial relationships between Race, Income, Education and the uninsured population I produced 24 choropleth maps to visualize the value of these variables in the state of Georgia on a census tract level. To accomplish this I downloaded a shapefile from Atlanta Regional Commission's Open Data Portal, of Georgia's 1959 census tracts. I joined a CSV table with the values for each of my variables to this shapefile based on tract ID's to create a spatial relationship between the ACS data for each census tract and the shapefile. The first two sets of maps I created were thematic maps for each variable. I did this for each variable from the 09-13 data as well as the 12-16 data. I used the same graduated color scheme for each variable with darker areas corresponding to higher percentages or values for each variable. The next set of maps uses the change value for each variable, and each census is either green or red. Green corresponds to an increase in the value of that specific variable while red indicates a decrease in that value. Finally, I did a hot-spot analysis in which tracts are symbolized by their z-scores in the dataset. This technique displays the quantiles for the range of values. I made 12 hot spot maps one for each IV and DV for 2009-2013, 2012-2016, change values. The next several pages highlight show these maps.

2009-2013 Variable
Thematic Maps



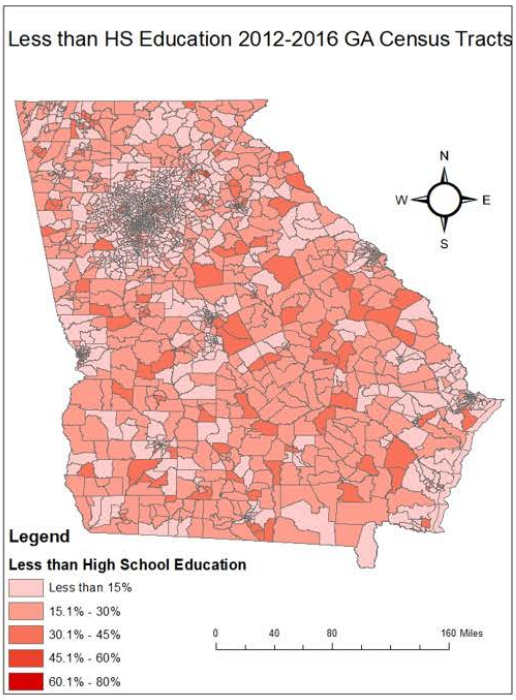




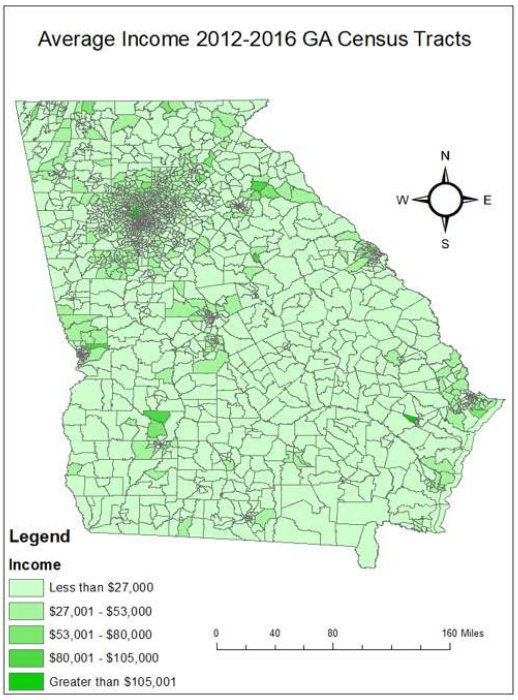


Source: American Community Survey

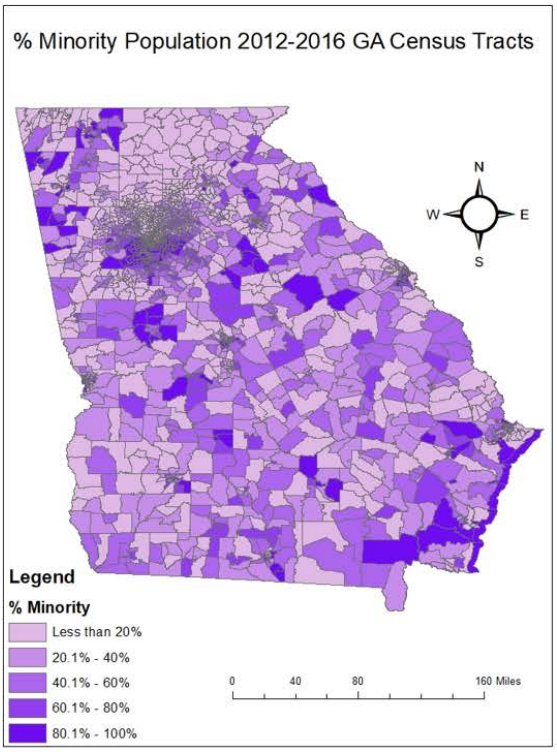
2012-2016 Variable Thematic Maps



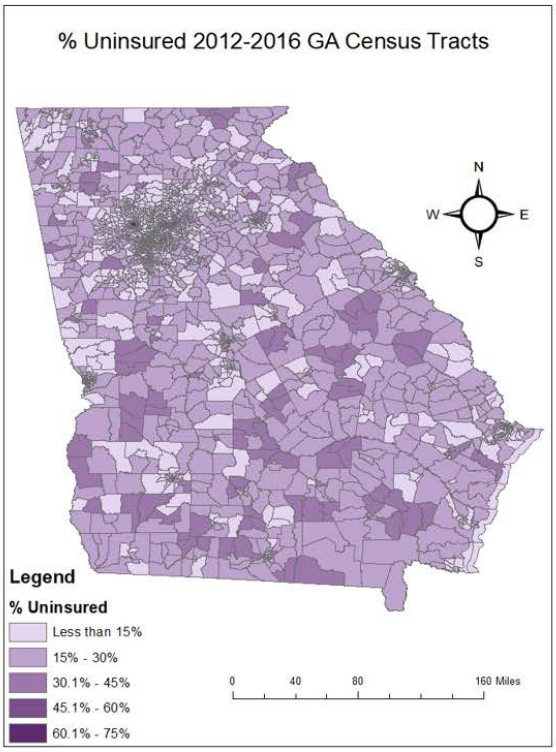
Source: American Community Survey



Source: American Community Survey

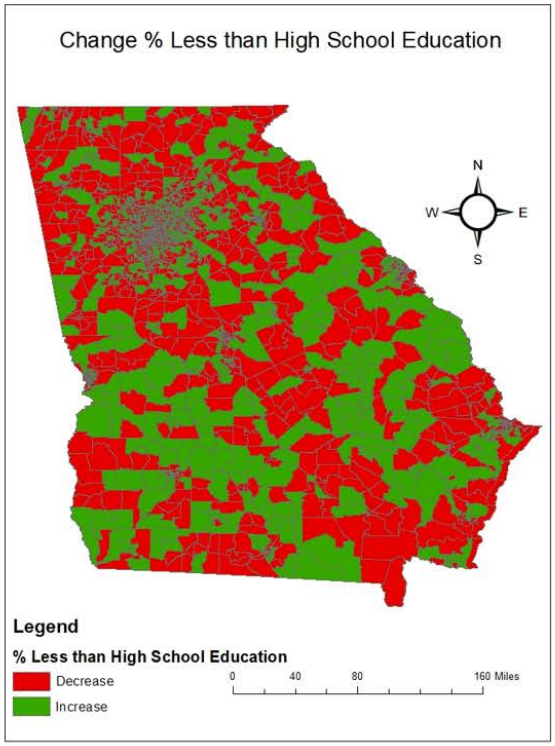


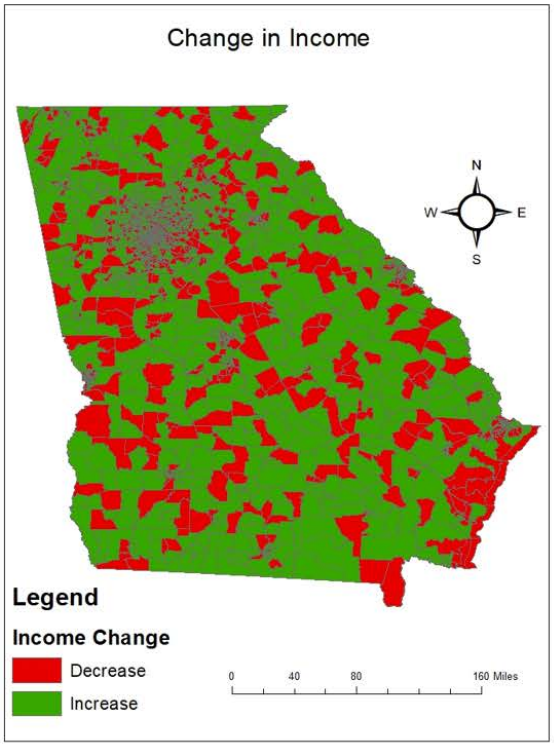
Source: American Community Survey

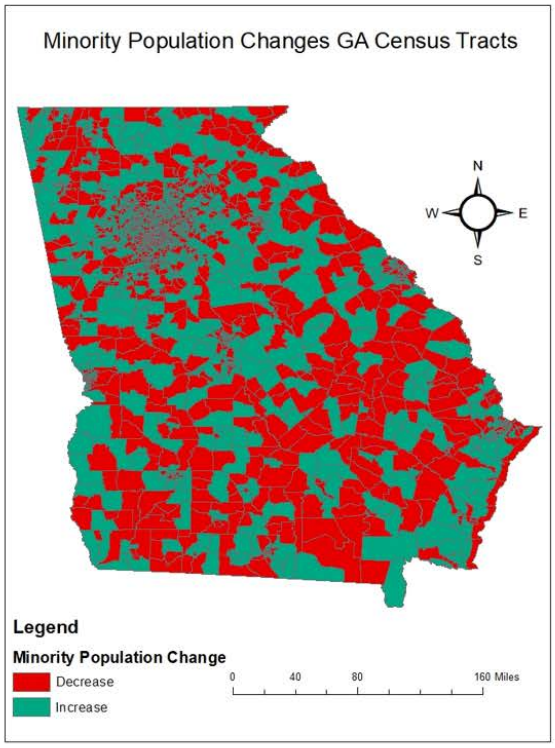


Source: American Community Survey

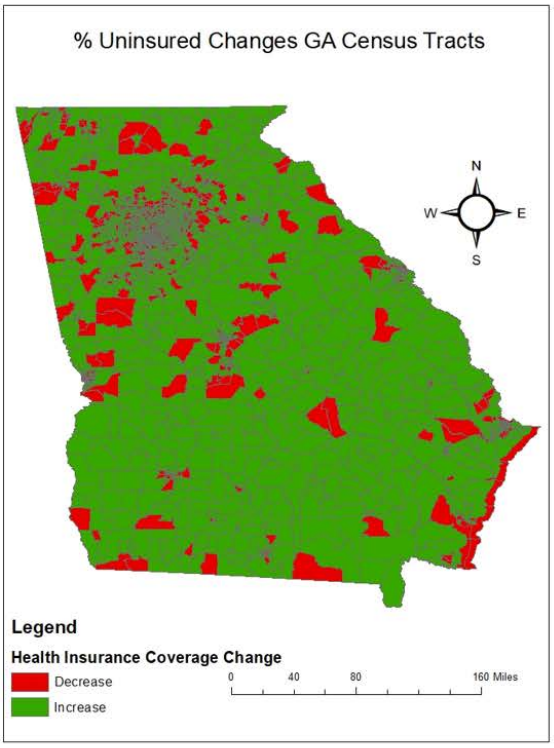
Change in Variable thematic Maps





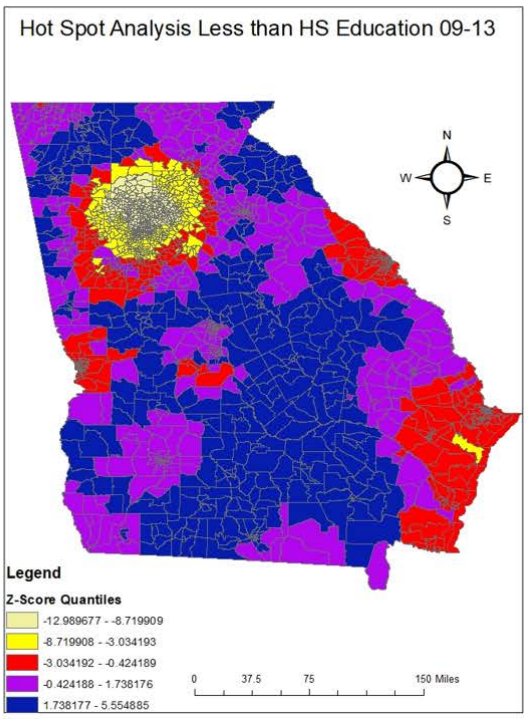


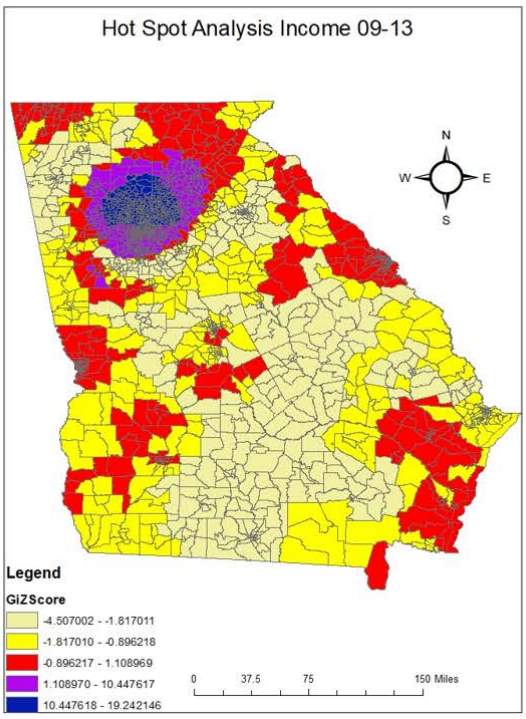
Source: American Community Survey

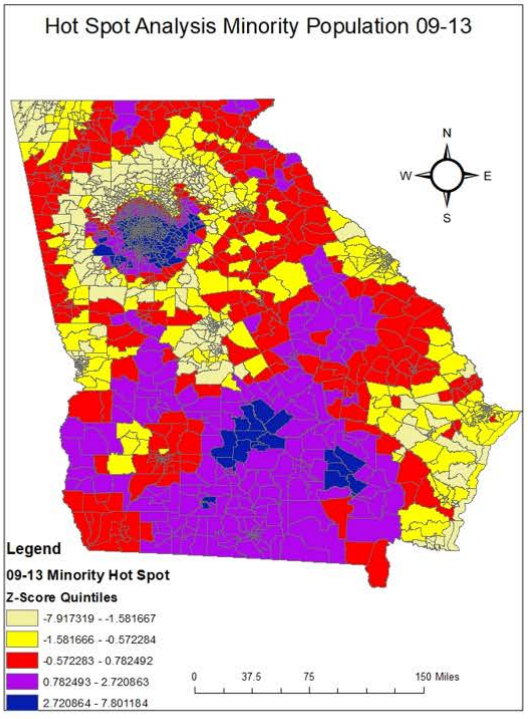


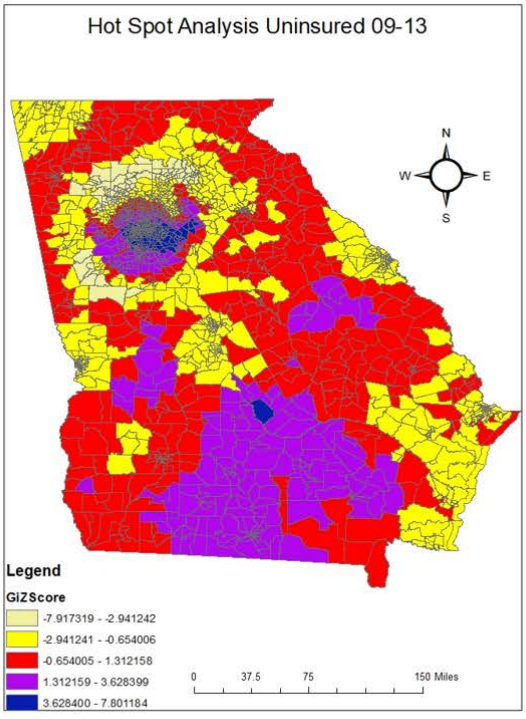
Source: American Community Survey

2009-2013 Hot Spot Analysis





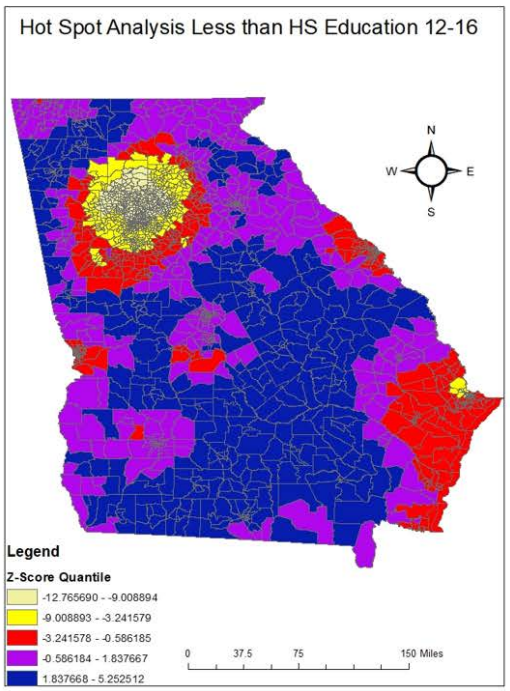


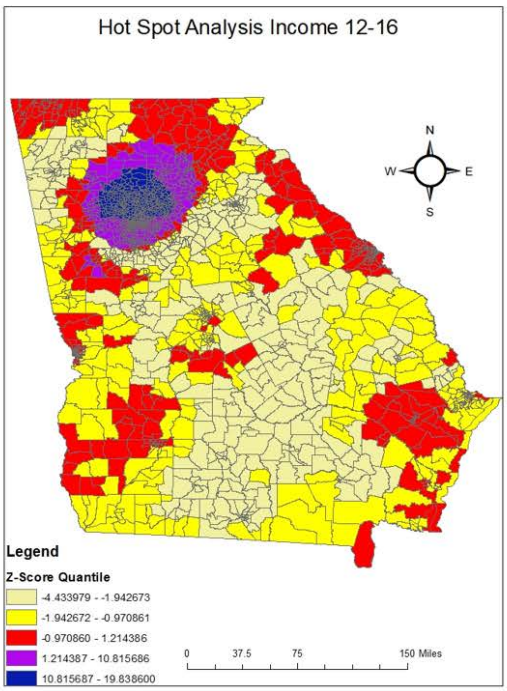


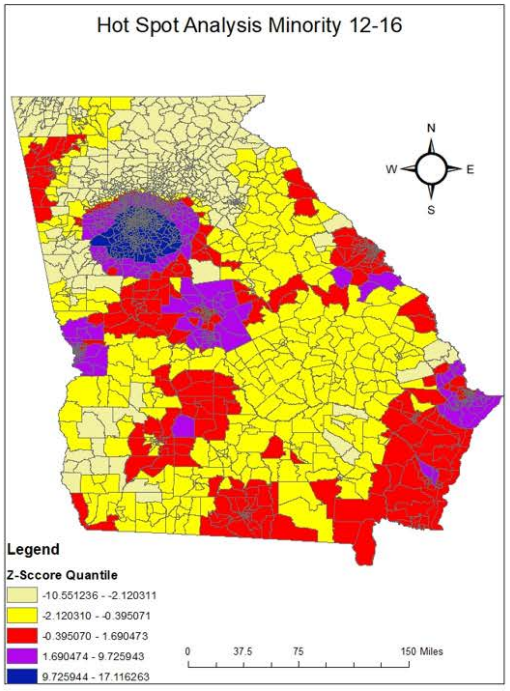
Source: American Community Surve

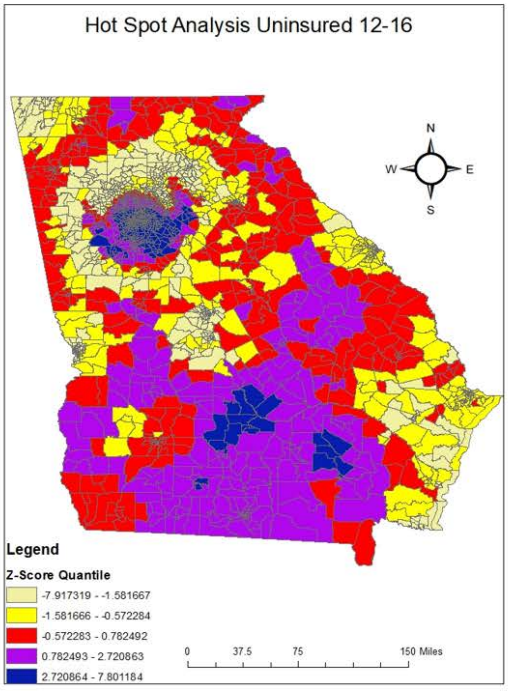
Source: American Community Survey

2012-2016 Hot Spot Analysis





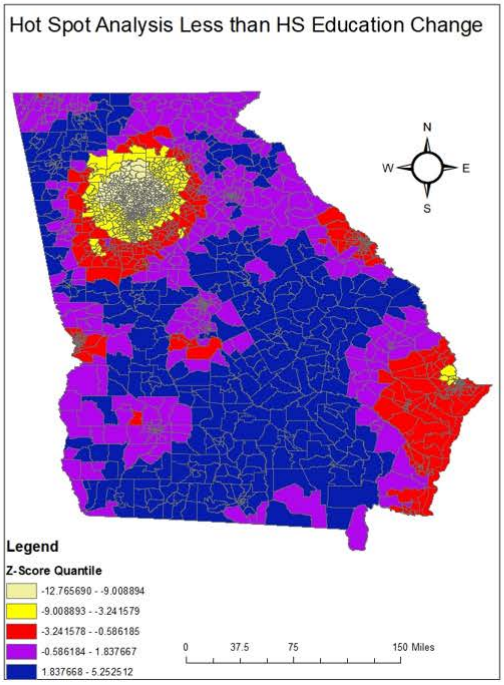


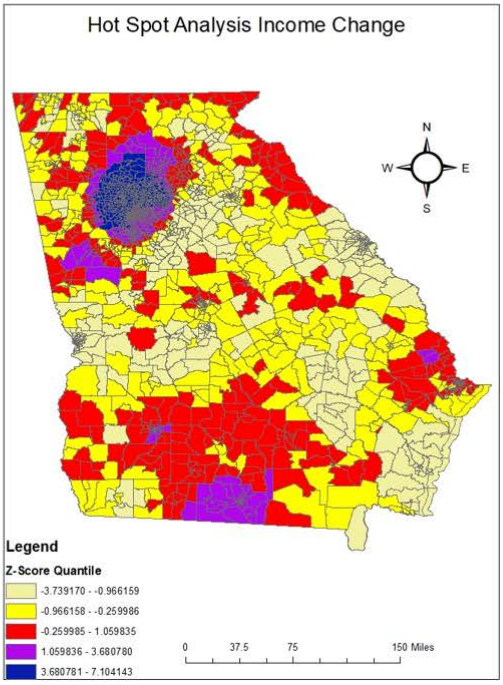


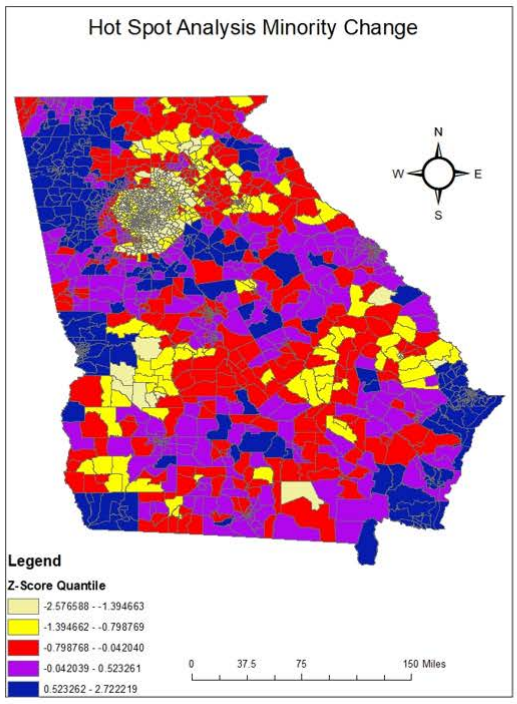
Source: American Community Survey

Source: American Community Survey

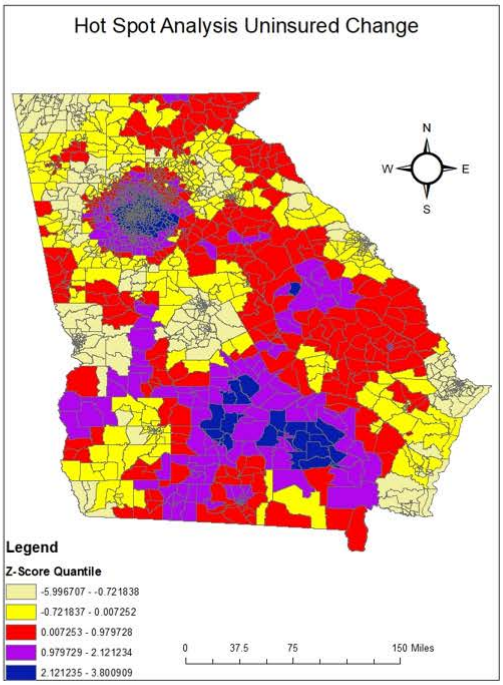
Change Hot Spot Analysis







Source: American Community Survey



Source: American Community Survey

Spatial Analysis Discussion

After an initial look at all the maps there are some patterns that emerge that span all of them. For instance, the Metro Atlanta area tracts tend to have higher per capita income, lower percentages with less than a high school education, lower uninsured rates, and a mix of high and low minority populations. These trends are the opposite for tracts in the south central and southeast portions of the state. There are not any starks contrasts between the 2009-2013 and 2012-2016 variable thematic maps meaning areas with either high or low value for each variable tended to not change significantly over the two time periods. The hot spot analysis for each variable between the two variables does not appear to change much either in terms of which area have high or low values for the IV's and DV. The change maps show generally positive trends for the state of Georgia. The most refreshing number was that 1540 of Georgia's census tracts had a decrease in the percentage of uninsured individuals. 1323 of the census tracts had an increase in income, and 838 of the census tracts had a decrease in the population with less than a high school education. Northwest Georgia had the greatest decrease in minority population. The hot spot analysis further quantifies these results. The hot spot change analysis shows that Metro Atlanta experienced significant differences in the two time periods for health insurance and income. There was a decrease in the uninsured population, increase in income. It was stagnant for the education and minority population variables. The south-central part of Georgia experienced a significant decrease in the percentage of population with less than a high school education, as well a decrease in the uninsured population, and the Northwest corner of Georgia showed the greatest decrease in minority population.

Significance and Conclusion

As evidenced by the previous literature the relationship between health insurance, health care, and other socio-economic factors is very complex and important in the US. This study aimed to quantify, analyze, and visualize those relationships with a focus on race, income, and education. I feel that the wide scope of the linear regression models explored several variations of these variables and their change over time. Time and time each model continued to show that these variables were significant and meaningful to each model. Unfortunately, these variables were not as good at predicting these values indicated by the low adjusted R-square and high AIC values for all of the models. For future research it would be useful to explore different metrics to quantify income, education, and race especially since they are all such complex factors.

The spatial analysis identified areas with low and high values for uninsured population, income, minority population, and education level. It also showed how these values had changed in certain areas between the two five-year time periods. Finally the hot-spot analysis showed where these changes were the greatest from the two data sets.

While lacking in many aspects, I feel that this project combining linear regression, and spatial analysis provided an in-depth, and robust exploration of the relationship between health insurance, race, income, and education in the state of Georgia. In the post-truth world that we live in today, it is important to be skeptical and rigorously pursue the truth in something as complex as this and I feel that this analysis moves the needle forward on this very important topic.

References

Molly Frean, Jonathan Gruber, Benjamin D. Sommers. Premium subsidies, the mandate, and Medicaid expansion: Coverage effects of the Affordable Care Act, Journal of Health Economics, Volume 53, 2017, Pages 72-86, ISSN 0167-6296, https://doi.org/10.1016/j.jhealeco.2017.02.004. (http://www.sciencedirect.com/science/article/pii/S0167629616302272)

Catherine Hoffman, Julia Paradise. *Health Insurance and Access to Health Care in the United States.* Ann N Y Acad Sci. 2008;1136:149-60. Epub 2007 Oct 22.

Joseph J. Sudano, David W. Baker. Explaining US racial/ethnic disparities in health declines and mortality in late middle age: The roles of socioeconomic status, health behaviors, and health insurance. Social Science & Medicine 62 (2006) 909–922.

National Health Expenditures Highlights 2016. Centers for Medicare and Medicaid Service. https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/Downloads/highlights.pdf.

ACS Information Guide. https://www.census.gov/content/dam/Census/programs-surveys/acs/about/ACS_Information_Guide.pdf.

https://www.r-project.org/about.html

R Documentation: Scale Function.
https://www.rdocumentation.org/packages/base/versions/3.5.1/topics/scale

Mean Centering Variables in SPSS.

https://www.spss-tutorials.com/mean-center-many-variables/

How to Interpret Regression Analysis Results: P-values and Coefficients (July 1st,2013)

http://blog.minitab.com/blog/adventures-in-statistics-2/how-to-interpret-regression-analysis-results-p-values-and-coefficients.