Analyzing Contributing Attributes to Homicides in the Southern United States

Jack Wibben – Geography 560 – Prof. Song Gao

Introduction

Crime is prevalent in every society, and the United States is no exception. As a person who is about to graduate and move into the real world, I want to know where I go is a safe place to live. The safety of a city can depend on a variety of different factors, and can be different in many different locations. It is impossible to know exactly how safe a neighborhood is, as random acts of violence can happen anywhere and anytime, but having a sense of how certain neighborhood characteristics contribute to levels of crime in a city can be a great way to predict where resources need to go to protect and prevent these things from happening, as well as finding the right place to live for the safety of your family and lifestyle. I hope with this data analysis to get a better idea on which variables have a positive or negative affect on the amount of homicides in a county. Due to socioeconomic tendencies in this country, I predict that areas of high unemployment and large populations would tend to have a larger percentage rate of homicides. This is my biased view before conducting any analysis, but I hope to learn more and have a better idea of the likelihood of these types of crimes after finishing this project.

Data

The data that I used for this project came from the GeoDa Data and Lab database, and from the specific dataset titled "South" (found in the link https://geodacenter.github.io/data-and-lab//county_election_2012_2016-variables/). I chose this dataset because it covers a very large area, giving the analysis a very broad area to analyze and, hopefully, giving us a better picture

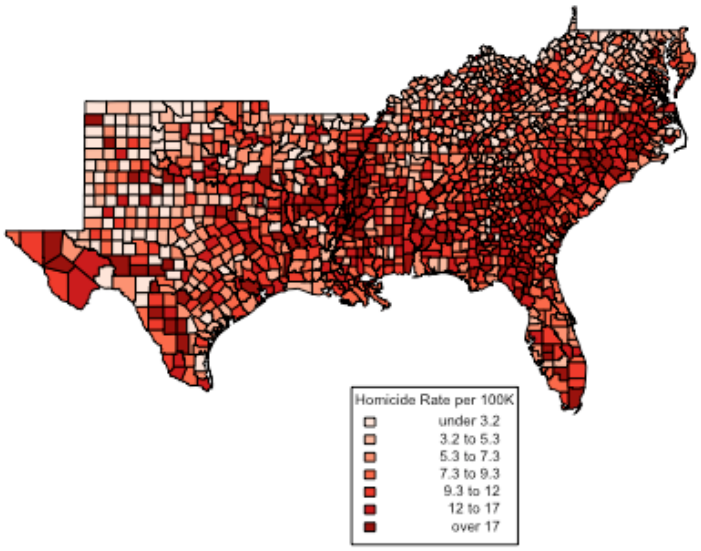
of variables that affect homicide rates in a given area. The broad data taken from many different states will help to avoid any localized bias found in any given area, and take data from both urban and rural counties. The data contains a total of 69 variables over 1,412 observations that covers data for four decennial years; 1960, 1970, 1980, and 1990. The large amount of observations at the county level gives the project a higher chance that the analysis comes back correct. The data is a bit old, so I condensed the project to only include the most recent observations, the ones from 1990. This left us with the variables PO90 (County Population in 1990), RD90 (Resource Deprivation in 1990), PS90 (Population Structure), UE90 (Unemployment Rate in 1990), DV90 (Divorce Rate in 1990), MA90 (Median Age in 1990), POL90 (Log of Population in 1990), DNL90 (Log of Population Density in 1990), MFIL89 (Log of Median Family Income in 1990), FP89 (Percent of Families Below the Poverty Line in 1989), BLK90 (Percent Black in 1990), GI89 (Gini Index of Family Income Inequality in 1989), and FH90 (Percent of Female Headed Households in 1990). These are the variables from the South dataset that I used in the analysis. The HR90 variable was used as the dependent variable, and the rest were used as independent variables to figure out which had the largest effect on the Homicide Rate in the states in the study, which are Texas, Oklahoma, Arkansas, Louisiana, Mississippi, Alabama, Georgia, Florida, North Carolina, South Carolina, Virginia, Maryland, West Virginia, Kentucky, and Tennessee.

Methods

The first part of the project is focused on finding the variables that have the most effect on the Homicide Rate per 100,000 residents in a given county. I utilized a Multiple Linear Regression test using the Homicide Rate per 100,000 residents in 1990 (HR90) as the dependent variable, while using the rest of the variables noted in the above Data section as independent variables as predicting variables. This was to conclude which variables have the greatest effect on/association with the homicide rate increasing or decreasing in the counties included in the dataset. After setting up and running the Multiple Linear Regression model, I got the Standard Regression Coefficients to determine which variables indeed had an effect, if any, on the homicide rate in these counties.

The second part of this project was a test to find out if there is any pattern to these variables spatially, especially that of homicide rates within the counties in the study. This was conducted through a Geographically Weighted Regression (GWR) Test, in which I used an Adaptive Kernel test using a fixed number of 5 nearest neighbors to set up the GWR test. A Goodness of Fit test and following R-Squared value will show us how strong the relationship between the GWR model and the response variable will be.

Results and Findings



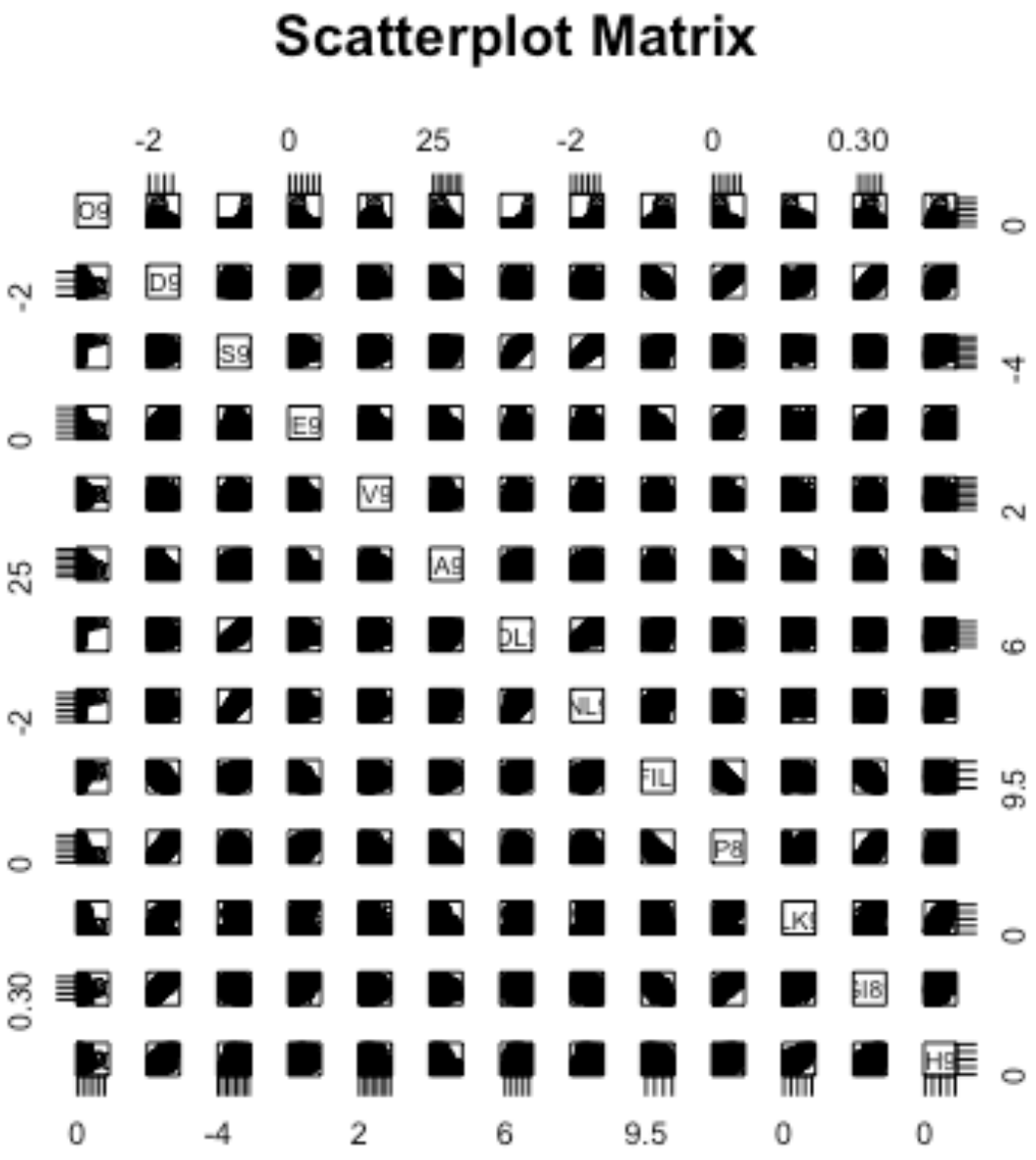
I first created a choropleth map giving us a nice look at where the homicide rates are prevalent in our study area before conducting any analysis on the data, which can be seen above.

For the first part of the test, running the Multiple Linear Regression model gave me an R-squared value of 0.3559. This shows that while the independent predictor variables did a decent job at explaining the rise/fall of homicide rate in these counties, but did not totally explain the dependent variable, and variation can be found from other variables omitted from the model. It also could explain that the model could be formatted better, and maybe a

Random Forest Regression Test would have produced results that would have been more accurate to the model. The final output statistics can be seen below:

```
Residual standard error: 5.671 on 1400 degrees of freedom
Multiple R-squared:  0.3559,    Adjusted R-squared:  0.3508
F-statistic: 70.33 on 11 and 1400 DF,  p-value: < 2.2e-16
```

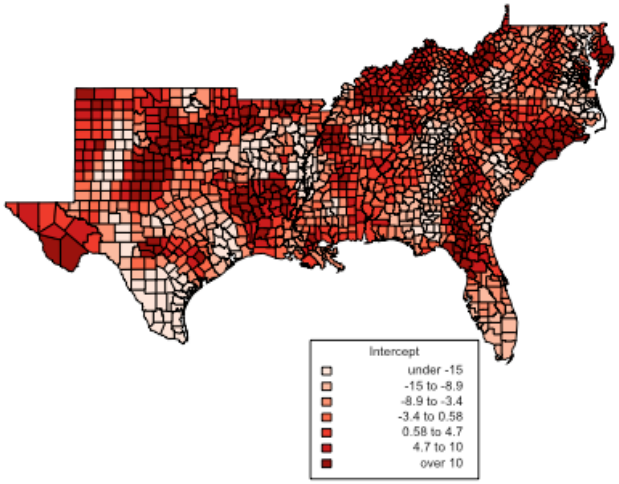
A pairwise scatterplot matrix of each variable was also conducted and can be seen in the image below, though with the amount of variables in consideration, it is rather tough to decipher:



The next step in this part of the test was getting the Standard Regression Coefficients for each of the independent variables, which allows us to see how each variable affects the dependent HR90 variable. The higher the absolute value of the value of each variable, the more effect this variable has on the variance of the HR90 dependent variable. From this test, I found that the Resource Deprivation variable (RD90) was miles ahead of the rest of variables, with a value of 1.475. The rest hovered around the 0.1 range, but Percentage of Families Below the Poverty Line (FP89), Log of Median Family Income (MFIL89) and Gini Index of Family Income Inequality (GI89) also had a rather large effect on homicide rate. The Log of Population Density (DNL90) and Percent of Female Headed Households (FH90) variables did not produce any results, and were represented with an NA. The full results can be seen below:

```
Standardized Coefficients::
(Intercept)        PO90        RD90        PS90        UE90        DV90        MA90       POL90       DNL90
 0.00000000  0.13125175  1.47537994 -0.15113195 -0.05117369  0.09386952  0.02811696  0.13342642          NA
    MFIL89        FP89       BLK90        GI89        FH90
 0.37828100 -0.30027468 -0.15863390 -0.30764172          NA
```

The second part of the test was the Geographically Weighted Regression test that determined whether there were any patterns in spatial deviance in homicide rate in the study area. Running the GWR test with 5 nearest neighbor indicators outputs a map that shows us clustering of high and low homicide rate (HR90) values in certain regions in the study area. The choropleth map can be seen below:



Determining how strong the GWR test is was found through a Goodness of Fit test, and following that, an R-Squared value that calculated the relationship. After running the code, I found an R-Squared value of 0.916 for the GWR test, which is very high and shows that the HR90 variable has a strong connection with spatial patterns in its variance.

Conclusions and Discussion

After running through the various analyses during the project, I came out with a few final conclusions. In determining the homicide rate per 100,000 people in a county, the variables that have the most effect were the resource deprivation, the log of median family income and the Gini Index for family income inequality. Income inequality is a serious problem in the United States, and it comes as no surprise to me that this is a large contributor to crime like homicide. As a situation gets tougher for a family, it can put them into dangerous places and sometimes crime can come with this. Also, we see from the GWR test that areas with higher homicide rates tend to be closer to each other, which also comes as not much of a surprise. This phenomenon is related with the first law of geography, which states that everything is related, but things closer to each other are more related than distant things. I believe that a study in such a large area can be a nice way to generalize a phenomenon like homicide, and the knowledge obtained here could be extrapolated to any area of the United States when considering variables that affect homicide.

Some extended research would need to be done to truly confirm the variables that affect homicide rate and other crimes. An R-Squared value of .35 is unfortunately pretty low, and if I were to do another project, I might try out other datasets in more confined spaces like a specific city, or try other regression tests like the Random Forests regression in order to further inform myself of the telling signs of an area being at risk for crimes like homicide. It also would be nice to find more recent information, because the most recent info for this dataset was in 1990. I believe this is new enough that it is still pertinent to how society works today in relation to homicide crimes, but a test on data in the 2010's would be even more likely to be an accurate teller for homicide rates and what affects them.

Acknowledgement

I want to thank Professor Song Gao for the assistance and guidance through the class material that prepared myself to conduct a project like this. This class has been a tough, but eye opening experiences to the types of research you can do with R and all of the data available online.

References

Homicides in the South (1960-1990). *GeoDa Data and Lab*. https://geodacenter.github.io/data-and-lab/south/ (last accessed 9 May 2019).

Executive Summary

Analyzing Contributing Attributes to Homicides in the Southern United States

Jack Wibben – Geography 560 – Prof. Song Gao

Introduction and Motivation

Crime is prevalent in every society, and the United States is no exception. As a person who is about to graduate and move into the real world, I want to know where I go is a safe place to live. The safety of a city can depend on a variety of different factors, and can be different in many different locations. I want to find out what are the driving factors towards high rates of crime in a given area, specifically the homicide rate.

Data Used

In order to conduct this research, I used the "South" dataset, obtained from the GeoDa Data and Lab website hosted by Github. The dataset had our dependent variable, Homicide Rate per 100,000 people (HR90), and a host of independent variables that were used in the study to determine the contributing factors towards a high value of HR90. All values were used from the year 1990 as it was the most recent study done. They are as follows:

PO90 (County Population in 1990), RD90 (Resource Deprivation in 1990), PS90 (Population Structure), UE90 (Unemployment Rate in 1990), DV90 (Divorce Rate in 1990), MA90 (Median Age in 1990), POL90 (Log of Population in 1990), DNL90 (Log of Population Density in 1990), MFIL89 (Log of Median Family Income in 1990), FP89 (Percent of Families Below the Poverty Line in 1989), BLK90 (Percent Black in 1990), GI89 (Gini Index of Family Income Inequality in 1989), and FH90 (Percent of Female Headed Households in 1990).

Methods

I used a Multiple Linear Regression test in order to find correlation between independent variables and the dependent variables, and this was determined from the Standardized Coefficient values. I also conducted a Geographically Weighted Regression test in order to find patterns in spatial variance of the dependent variable.

Results and Conclusion

From the analysis, I found that the variables that contributed most the homicide rate were resource deprivation, the log of median family income and the Gini Index for family income inequality. This shows that how money is distributed amongst a population is a definite determinant towards crime levels, and the more disadvantaged an area, the more likely homicide is a common occurrence. Also, I found that areas with higher homicide rates tend to be closer to one another, with a high R-Squared value in the GWR testing. Testing more variables in additional analysis could prove helpful, and would be the main focus of additional work on a project like this.