**A Spatial Analysis of Predictors of Different Types**

**of Crime in Chicago Community Areas**

May 13th, 2014

Brett Beardsley

brett.a.beardsley@gmail.com

MGIS Capstone Project

The Pennsylvania State University

Faculty Advisor:

Dr. Stephen A. Matthews

# Abstract

Chicago has more than twice the crime rates of Los Angeles and New York City for almost all major crime types. In this paper I examine eight crime types and salient socioeconomic predictors in Chicago based on exploratory spatial data analysis (ESDA) and spatial econometric approaches using *Open Geoda*. I use the FBI Uniform Crime Report (2007-2011) and 5-year estimates from the American Community Survey (ACS) for the 77 Chicago Community Areas for the crime and socioeconomic data respectively. The analytical approach included ESDA based on mapping and descriptive statistical analysis of each crime type and potential salient predictors. Spatial weights files based on Queen's first-order connectivity metrics were used to calculate both global spatial autocorrelation measures (i.e., Moran's I). OLS regression models without spatial weights served as a baseline model. These baseline models were repeated with spatial weights and spatial diagnostics checked. As necessary, models were re-specified based on either spatial lag or spatial error model structures. My goals are to indentify the variables associated with overall crime and specific crime types in Chicago, measuring and controlling for spatial effects, and to identify any common patterns and relationships among the statistically significant variables across crime type.

Contents

[Abstract 2](#_Toc387696825)

[Problem Description 4](#_Toc387696826)

[Research Questions: 5](#_Toc387696827)

[Literature Review 5](#_Toc387696828)

[Chicago studies 6](#_Toc387696829)

[Missing information 9](#_Toc387696830)

[Approach/methods 10](#_Toc387696831)

[Download and Process Crime and Census Data 13](#_Toc387696832)

[Data Analysis 15](#_Toc387696833)

[Descriptive Statistics 18](#_Toc387696834)

[Correlation Matrices 21](#_Toc387696836)

[Regression 24](#_Toc387696837)

[OLS Regression-Total Crime Rate 25](#_Toc387696838)

[Spatial Lag Model-Total Crime Rate 28](#_Toc387696839)

[OLS Regression-Violent Crime Rate 30](#_Toc387696840)

[Spatial Error Model-Violent Crime Rate 32](#_Toc387696841)

[OLS Regression-Property Crime Rate 33](#_Toc387696842)

[Spatial Lag Model-Violent Crime Rate 35](#_Toc387696843)

[Results and Conclusions 37](#_Toc387696844)

[Limitations 38](#_Toc387696845)

[Further Studies 38](#_Toc387696846)

[Works Cited 40](#_Toc387696847)

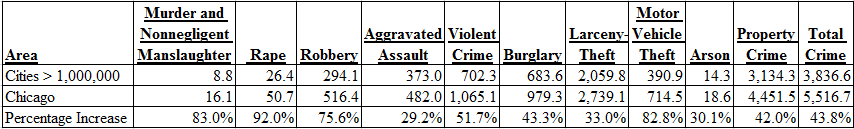
[Appendix A: Selected Choropleth Maps of Variables and Total Crime Rate Chart 43](#_Toc387696865)

[Appendix B: Total Crime Rate Chart and Correlation Matrices 45](#_Toc387696866)

# Problem Description

Chicago is the third most populous city in the United States with a population of 2,714,856. Many large cities have their fair share of problems with crime and Chicago is not immune to these problems. In recent years Chicago has gained worldwide notoriety for the large number of murders and shootings that have occurred in the city. The violence is so pervasive that it inspired one foreign film maker to create a documentary named "Chi Raq" (Huffington, 2013). In fact from 2001 to May of 2012 there were 3,080 more people murdered in Chicago than U.S. troops killed in Afghanistan (Peterson, 2012). Although the murders and shootings may get most of the headlines, the statistics also tell a story of other crimes plaguing Chicago. When examining the crime rates for the eight crimes in the Federal Bureau of Investigation's (FBI) uniform crime report, which are murder and non negligent manslaughter, rape, robbery, aggravated assault, burglary, larceny-theft, motor vehicle theft, and arson, in Chicago and in the nine cities (including Chicago) in the United States with a population of at least 1 million it is clear that Chicago has a problem with crime. The crime rates from 2011 is summarized in Table 1 below.

**Table 1: 2011 FBI Uniform Crime Report Rates per 100,000 People (Federal Bureau of Investigation, 2011 and City of Chicago, 2014)**



You can see that crime rates in Chicago were higher for every crime from the FBI's Uniform Crime report in 2011 than the average rate in cities in the United States with a population greater than 1 million. The rates for violent crime combined all murder and non negligent homicides, rapes, aggravated assaults, and robberies and the property crime rates included all burglaries, larceny-thefts, motor vehicle thefts, and arsons. The percentage increases for Chicago compared to cities with populations over 1 million range from 92% for rape to 29% for aggravated assault. The crime that has been given the most attention, murder and non negligent manslaughter, is not the highest percentage wise above the average for all cities over 1 million. The fact that every crime has a least a 25% higher rate in Chicago than in other cities in the United States with over 1 million people is alarming and should be examined.

# Research Questions:

Through this study I hope to answer numerous questions. Chicago is divided into 77 Community Areas and most previous spatial analyses on crime in Chicago used census tracts or blocks as the geographic unit of measurement. Census blocks and tracts mean little to nothing to most people. One question I want to answer is what Chicago Community Areas had the highest rates of murder and non negligent manslaughter, rape, robbery, aggravated assault, burglary, larceny-theft, motor vehicle theft, arson, violent crime, property crime, and total crime from 2007 to 2011. Next I will attempt to identify the predictors associated with overall, property, and violent crime in Chicago Community Areas during that time period. I would also like to identify the most influential predictors for each crime type. I will also identify any patterns and relationships among the statistically significant predictors across crime type.

# Literature Review

There has been a very large amount of literature on crime and the various factors that are related to it from a spatial perspective. From the research I have conducted I have found writings about the geographic study of crime going as far back as 1929 (Shaw, 1929). There have been many different factors examined in these studies using various methods on different forms of crime and they have taken place all over the world. In 1932 White plotted felonies across zones and on a semi-logarithmic scale in Indianapolis and calculated their correlation coefficients in an effort to see the effect social factors had on the crimes and if there were any visible spatial patterns. In another study Boggs (1965) examined the relationship of crime rates of different types of crime at different land use areas and when possible the location of the perpetrator by using factor analysis and correlation in St. Louis. In more recent times geographic information systems (GIS) has become a useful tool for spatial analysis of crime. GIS has been used to explore crime and the distribution of disadvantaged, middle class, and affluent neighborhoods in Merseyside County, England (Bowers & Hirschfield, 1999). A study in Sao Paulo, Brazil analyzed temporal (weekend, pay day, day time) and weather factors on homicide rates and homicide hot spots using Ordinary Least Squares (OLS) regression (Ceccato, 2005). A more recent study involved using exploratory spatial data analysis tools and Bayesian models to examine spatial patterning and predictors of property crime in Seattle, Washington (Matthews, Yang, Hayslett, & Ruback 2010). As you can you can see geographic analysis of crime has a rich and varying history. A common finding in these studies is that crime is spatially clustered and a small percent of places account for a large percentage of the crime in a given place. I will now focus mainly on the analysis of crime in Chicago from a selection of papers in the past 15 years.

## Chicago studies

Earls, Morenoff, and Sampson (1999) wanted to examine the structural sources and spatially embedded nature of three mechanisms that produce collective efficacy for children. They surveyed 8,782 Chicago residents in 1995 to examine variations in intergenerational closure, reciprocal local exchange, and shared expectations for informal social control across 342 neighborhood clusters from the 1995 Project on Human Development in Chicago Neighborhoods (PHDCN). Intergenerational closure and reciprocal local exchange were measured using a five item scale and shared expectations for social control was measured using a three item scale. The authors used hierarchical linear models to perform their analysis. The results of their analysis showed that concentrated affluence, low population density, and residential stability are the most consistent predictors of good levels of reciprocal local exchange and intergenerational closure. They also showed numerous times throughout their study how Latinos and especially blacks are at an automatic disadvantage compared to whites in terms of efficacy for their children. This condition was consistent across almost all situations and lead them to look for answers as to why that is. They also found that collective efficacy of children in surrounding neighborhoods has a direct relationship with a given neighborhood's internal collective efficacy. In sum they found that minorities are at a disadvantage to whites almost automatically and we must look at factors within neighborhoods to explain some spatial patterns. They also stressed that not every variation can be explained (Earls, Morenoff, & Sampson, 1999).

Morenoff's (2003) study addresses two questions by conducting a study in Chicago: what are the mechanisms through which neighborhood structural composition is related to health and are influences on health limited to the immediate neighborhood or do they extend to a wider geographic context. To conduct the study the author used data from the 1990 census, the 1995 Chicago Police Crime Statistics, and the 1995 PHDCN. The neighborhoods for the study were defined as the sampling units for the PHDCN Community Survey. The author used 29 variables that were broken up into five categories or groups. The categories were neighborhood context and then four individual-level variable groups, maternal race, ethnicity, and nativity, maternal sociodemographic characteristics, maternal health behavior, and maternal biomedical measures. These variables effects on continuous birth weight and dichotomous low birth weight were analyzed using hierarchical modeling techniques and the local Moran statistic. The results showed that violent crime and the combined scale of reciprocal exchange and participation in voluntary associations are the two most robust predictors on birth weight. He notes that neighborhood structural factors do not matter for the health of a baby, but lead to conditions which lead to stress and violence, which in turn effect the health of a baby. Morenoff also found that unobserved factors must be having an effect on low birth weight. This led him to conclude that factors outside of the neighborhood being examined may also influence birth weight there. In sum Morenoff stresses that we must look at the effect that surrounding neighborhoods have on one another (Morenoff, 2003).

Raudenbush, Sampson, and Sharkey (2008) examined verbal ability among African American children in Chicago. Their goal was to try to identify sources that effected cognitive ability throughout a person's childhood. They focused on African-Americans because poverty in African-American communities is overwhelming compared to whites and Latinos. They obtained their data from interviewing 6,234 African-American children from ages 3 to 18 at intervals from 1995 to 2002. The children were randomly sampled from 80 neighborhood clusters. The children were followed across the country to wherever they moved to continue the interviews in future years. The interviews were used to determine verbal cognitive ability. Census data from 1990 and 2000 was used to measure six variables that determined a concentrated disadvantage measure. Lastly they also looked at 21 time invariant and time-varying covariates. The authors used marginal structural models and Inverse Probability of Treatment Weighting (IPTW) to do their analysis. They found that concentrated disadvantage can have a huge effect on their verbal ability. It can have effects that are equal to taking away years of schooling and these effects can linger even after a child leaves a severely disadvantaged neighborhood. In sum the authors make a point that schools can only do so much to help children that live in severely disadvantaged neighborhood. This shows that policy makers should concentrate on improving poor neighborhoods as much as improving poor schools (Raudenbush, Sampson, & Sharkey, 2008).

Graif and Sampson (2009) examined the connection of immigration and diversity to homicide through geographically weighted regression. They also wanted to examine spatial heterogeneity across neighborhoods. The following variables were used from the 1990 and 2000 census: population density, concentrated disadvantage, unemployment percentage, poverty percentage, female-headed households, public assistance households, residential stability, and language diversity. These variables were measured across census tracts in Chicago. The dependent variable was homicide rate per 100,000 residents. They found that socioeconomic disadvantage increases homicide rate and residential stability decreases homicide rate. Another finding was that overall immigration had an insignificant role in both increasing and decreasing homicide rates. There was evidence in some areas that immigration concentration decreased homicide rates, which shows support for immigrant neighborhoods being able to enforce norms and keep their families intact. This in turn will help decrease the violence. They also found that increased diversity or cultural heterogeneity decreases the risk of homicide. Based on these findings the authors final message is that global models of neighborhood crime often do not look deep enough at the many factors that can shape crime (Graif & Sampson, 2009).

Berg, Brunson, Stewart, and Stimons (2011) examined the role cultural heterogeneity plays in shaping adolescent decision making concerning violence in poor neighborhoods. The study was conducted on metropolitan census tracts in Iowa and Georgia and on African-American children in two waves. The first wave was in 1997 and involved 10-13 year olds and the second wave was in 1999 with the same children. Roughly 800 children were interviewed each time. The dependent variable was violent delinquency and they used 17 independent neighborhood-level variables. The authors used multilevel modeling techniques to conduct their analysis. They found that cultural heterogeneity and concentrated disadvantage are associated with youth violence. They also found that the negative effect of individual conduct frames is less pronounced in neighborhoods where there is greater cultural heterogeneity. The conclusions they drew from this analysis is that socioeconomic disadvantage fosters non conventional models of behavior when dealing with violence (Berg, Brunson, Stewart, & Stimons, 2011). This means that children see conflicting examples of how acceptable violence is in certain situations in areas with greater socioeconomic disadvantage. This causes them to grow up accepting violence as an answer more often than children who grow up in areas with less cultural heterogeneity where violence is almost always unacceptable. Unlike children in areas with greater cultural heterogeneity, this teaches children to almost never resort to violence. Lastly the authors suggest that future work should attempt to look at mechanisms within neighborhoods to measure variation in the strength of cultural processes on behavior. In sum the authors found that the way that a child's neighborhood model behavior concerning violence has a huge impact with how they will deal with violence in their own lives (Berg, Brunson, Stewart, & Stimons, 2011).

Arnio and Baumer (2012) examined the role of the demographic factors on neighborhood crime rates and explored the spatial heterogeneity across Chicago census tracts from 2005 to 2009. The variables they examined to see their effect on crime rates were change in logged REO foreclosure rates (2007-2009), residential stability, immigrant concentration, socioeconomic disadvantage, percent black, population size and density, percent divorced, pre-existing vacancy rate, and logged REO foreclosure rate (2007). The crimes they examined were restricted to robbery, burglary, and homicide and they used Chicago census tracts as their area of measurement. They used Maximum Likelihood and geographically weighted regression models to perform their analysis. They found that there was significant local variations for the relationship between percent black and immigrant concentration and burglary and robbery rates. There was also significant local variation in the relationship between socioeconomic disadvantage and robbery rates and the relationship between residential stability and burglary rates. Lastly there was also significant local variation in the relationships between change in logged REO foreclosures and robbery and burglary (Arnio & Baumer, 2012). In sum the article showed why it is useful and important to examine changes in demographics and crime in the areas surrounding the focus area. It proved that looking at crime and demographics globally is not always the best approach. You can see a summary of the Chicago studies in Table 2 below.

**Table 2: Summary of Literature Review of Spatial Analysis on Crime in Chicago**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Earls, Morenoff, and Sampson | Morenoff | Raudenbush, Sampson, and Sharkey | Graif and Sampson | Berg, Brunson, Stewart, and Stimons | Arnio and Baumer |
| Predictors | -Intergenerational Closure  -Reciprocal Local Exchange  -Shared Expectations | -Neighborhood Context, Maternal Race, Ethnicity, and Nativity  -Maternal Sociodemographic Characteristics  -Maternal Health Behavior  -Maternal Biomedical Measures | -Concentrated Disadvantage -Age of Subject and Caregiver  -Sex of Subject and Caregiver  -U.S. Citizenship of Caregiver  -Caregiver's Educational Attainment  -Family Criminality  -Caregiver's Domestic Violence Index  -Caregiver's Major Depression  -Caregiver's Social Support -Caregiver's Total Household Income  -Caregiver's and Spouse or Partner's Employment Status  -Caregiver's Occupational Status  -Caregiver's Socioeconomic Status  -Home Ownership  -Household Size  -Caregiver's Marital Status | -Population Density  -Concentrated Disadvantage  -Percentage  Unemployed  -Poverty Percentage  -Female-Headed Households  -Public Assistance Households  -Residential  Stability  -Language Diversity | -Cultural Heterogeneity  -Concentrated Disadvantage  -Neighborhood Social Ties  -Homicide Rate  -Conduct Frames | -Change in Logged REO Foreclosure Rates (2007-2009)  -Residential Stability  -Immigrant Concentration  -Socioeconomic Disadvantage  -Percent Black  -Population Size  -Population Density  -Percent Divorced  -Pre-existing Vacancy Rate  -Logged REO Foreclosure Rate |
| Outcome Variables | Collective Efficacy for Children | Birth Weight | Verbal Cognitive Ability | Homicide Rate per 100,000 Residents | Violent Delinquency | -Robbery and Homicide Rate per 100,000 Residents  -Burglary Rate per 100,000 Housing Units |
| Time Period and Place | 1995-Chicago | 1990 and 1995-Chicago | 1995-2002-Chicago and Throughout the U.S. | 1990 and 2000-Chicago | 1997 and 1999-Georgia and Iowa | 2005-2009-Chicago |
| Age and Number of Subjects | 18 and Older -8,782 | Women of Child Bearing Age-101,662 | 3 to 18 Years Old-6,234 African-Americans | All Ages | 10 to 15 Years Old-800 African-Americans | All Ages |
| Unit of Analysis | 342 Neighborhood Clusters | 343 Neighborhood Clusters | 80 Neighborhood Clusters | Census Tracts > 100 Residents | 106 Census Tracts | Census Tracts > 100 Residents |
| Method of Analysis | Hierarchical Linear Models | Hierarchical Modeling Techniques and the Local Moran Statistic | Marginal Structural Models and Inverse Probability of Treatment Weighting | Geog. Weighted Regression | Multilevel Modeling Techniques | Maximum Likelihood and geographically weighted regression models |

## Missing information

None of the articles covered in the literature review used Chicago Community Areas as their unit of analysis. They all used census tracts or neighborhood clusters which are made of census tracts because that unit of analysis provided a much larger sample size that is more ideal for many more kinds of analysis. As previously mentioned there are only 77 Community Areas in Chicago, which is a small sample size for certain types of analysis. However, a study using the official Chicago Community Areas would clearly be valuable as those are places people recognize. This is because almost no one knows what tract they live in and almost everyone knows what Community Area they live in. The outcome variables of the Chicago studies also did not include as many types of crime as I wanted to include. The Chicago studies were also slightly out of date with the most recent one going through 2009, so my study should give a more up to date view on the spatial analysis of predictors of crime in Chicago.

# Approach/methods

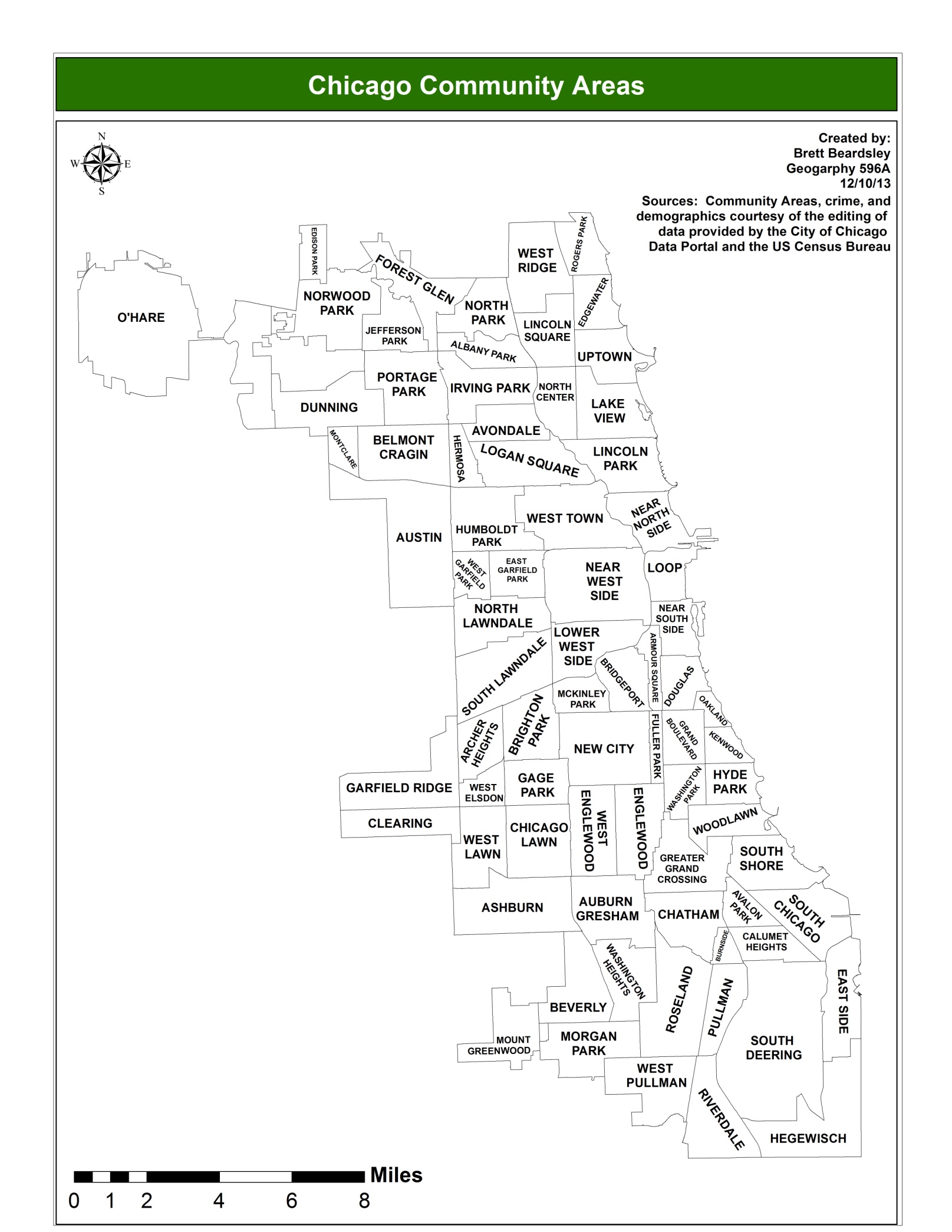
The preliminary steps for this project involved choosing my outcome variables, predictor variables, and the unit of analysis. General research about crime and the literature review helped greatly with choosing the outcome and predictor variables. Through internet research it was quickly discovered that the FBI Uniform Crime Report crimes were almost universally used throughout the United States. The FBI's site also had the eight crimes of murder and non negligent manslaughter, rape, robbery, aggravated assault, burglary, larceny-theft, motor vehicle theft, and arson broken down into violent crimes (murder and non negligent manslaughter, rape, robbery, and aggravated assault) and property crimes (burglary, larceny-theft, motor vehicle theft, and arson). Total crime (violent crime plus property crime) was also added as an outcome variable for a total of 11 outcome variables. It is also important to note that all crimes would be converted to a rate of number of crimes per 100,000 people.

The outcome variables were chosen based on findings from the literature review and on the author and his advisor wanting to include variables that they believed would have a relationship with crime. There was also a quick look given to Census website and Chicago's Data Portal to determine what statistics were available. Percent of household below poverty, percent of African-American, and percent foreign born were all found to have significant relationships with crime in studies in the literature review. Percent aged 16 and over unemployed, percent aged 25 and over without a high school diploma, per capita income, percent female-headed households, percent white, population density (population per square mile), and percent foreign born were included in studies in the literature review. Percent overcrowding, percent aged under 18 or over 64, percent of the population 15 to 24 years old, percent of households built before 1960, the median value of owner occupied housing units, percent of housing units renter occupied, and percent of households with three or more bedrooms were added at the author and his advisor's knowledge on crime and its causes. They were more exploratory than the others predictors. The idea was that they may prove valuable although they were not included in the studies in the literature review. Percent overcrowding, percent of the population between 15 and 24 years old, percent of households built before 1960, and percent of housing units renter occupied were expected to increase crime rate as each predictor increased. Percent aged under 18 or over 64, median value of owner occupied housing units, and percent of households with three or more bedrooms were expected to decrease crime rates as each predictor increased. Table 3 summarizes the outcome and predictor variables.

**Table 3: Outcome and Predictor Variables for Spatial Analysis of Crime in Chicago**

|  |  |
| --- | --- |
| Outcome Variables (rates per 100,000 people) | Predictor Variables |
| Murder and Non Negligent Manslaughter Rate | Percent Overcrowded Housing |
| Rape Rate | Percent Households Below Poverty |
| Robbery Rate | Percent Aged Over 16 and Unemployed |
| Aggravated Assault Rate | Percent Aged Over 25 Without a High School Diploma |
| Burglary Rate | Percent Aged Under 18 or Over 64 |
| Larceny/Theft Rate | Per Capita Income |
| Motor Vehicle Theft Rate | Percent Female-Headed Households |
| Arson Rate | Percent African American |
| Violent Crime Rate | Percent White |
| Property Crime Rate | Percent Population Aged 15 to 24 |
| Total Crime Rate | Percent Households Built Before 1960 |
|  | Median Value of Owner Occupied Housing Units |
|  | Population Density (Population per Square Mile) |
|  | Percent Vacant Housing Units |
|  | Percent Foreign Born |
|  | Percent Housing Units Renter Occupied |
|  | Percent Housing Units with Three or More Bedrooms |

The unit of analysis for this project changed numerous times between Chicago Community Area and census tract. As previously mentioned census tracts would have been the more conventional unit of analysis here as there would have been 805 census tracts used whereas there are only 77 Community Areas. There was also some thought given to aggregating census tracts into larger areas similar to what was done in the literature review studies that used neighborhood clusters. After much thought it was determined that the unit of analysis would be Chicago Community Areas for several reasons. One was that it would distinguish my study on spatial analysis of crime in Chicago from the vast majority of similar studies. Another reason is that spatial regression can be performed on a sample of 77. Finally, as was previously mentioned almost no one knows what census tract they live in and almost everyone knows what Community Area they live. Therefore the results of a study using Chicago Community Areas would be more valuable to most people than a study using census tracts. Figure 1 shows a labeled map of Chicago Community Areas.



**Figure 1 : Chicago Community Areas**

## Download and Process Crime and Census Data

All crime data for the eight crimes from 2001 to 2012 from the Uniform Crime report were downloaded from the Chicago Data Portal. Each crime dataset had to be downloaded separately because the database was so large. Once downloaded the data was in excel sheets in the form of point data for each crime and had the date and Community Area already in it, which was extraordinarily convenient for my purposes. It is important to note that the Chicago Police Department considers more forms of sexual assault as "rape" than the FBI (Goode, 2011). For this study any criminal sexual assault in Chicago was considered rape. This results in a larger number of "rapes" in Chicago in this study than if the FBI's definition was used. Once all of the crime data was collected it was time to examine the possible sources for the predictor variables

of interest.

As was previously mentioned it was known that each predictor variable was available from the Census website or the Chicago Data Portal. The Chicago Data Portal's data was once again conveniently packaged for me by community area. It was taken from the 5-year ACS from 2007-2011. It is important to note that the ACS is not a complete census, but a survey of a small proportion of the population every year (Census, 2014; What). The six predictor variables found through the Chicago Data Portal were percent overcrowded housing, percent of households below poverty, percent aged over 16 and unemployed, percent aged over 25 without a high school diploma, percent aged under 18 or over 64, and per capita income. I did not need to manipulate the data because it was available to download in Excel by Community Area. Once it was known that the first six predictors were from the 5 year ACS from 2007 to 2011 it was strongly preferred that the other eleven predictors came from the same data set. Before that decision was finalized I wanted to ensure that the 2007-2011 ACS was the most up to date data available. Through exploring the Census's website it was confirmed that the 5 year ACS data from 2007 to 2011 was indeed the most up to date data. It is also important to note that this research was taking place in November 2013 and the 5 year ACS data from 2008 to 2012 was not available yet. Now came the large task of processing and manipulating all of the data I had into a form I needed it to be in.

First, the crime data was cut down from 2001 to 2012 to 2007 to 2011. It was then added into ArcMap where the points were spatially joined to a layer of the Community Areas. The joined Community Areas now had summary statistics of how many times each type of crime had taken place within their boundaries from 2007 to 2011. All that was left to do in terms of the crime data was to convert the sums into rates and that would happen once the population data was joined to the shapefile. The six predictors were already in the form of percentages or per capita income from the Chicago Data Portal. This was perfect and all that was needed for them was to be brought into ArcMap and joined to the shapefile with the crime statistics. Now came the extraordinarily time consuming part of processing and manipulating the eleven remaining variables from the ACS.

The ACS data came from a table of 2,035 attributes which were coded. The codes all corresponded to an attribute that was found on a key. Since the data was so large with these attributes it had to be minimized in the tables to include only what was needed. This was very painstaking work because of how many attributes had to be sifted through. The remaining eleven predictors also came in the form of totals, so I had to keep the total population they were being compared to in the table. For some predictors such as population between the ages of 15 and 24 I had to collect up to eight attributes for the different statistics that would be compiled into the predictor I wanted. The ACS data was also in the form of census tracts, which was a problem because that was not the unit of analysis I was using. Luckily Chicago Community Areas are almost exactly made up of census tracts, so it was possible to sum all of the data from each tract into a total for a Community Area. This took a large amount of effort though. The ACS data was brought into ArcMap with a Community Area shapefile and centroids for each Community Area were created. The ACS tables were given a Community Area name based on which Community Area's centroid was closest to a given census tract. Unfortunately that only properly identified approximately 650 out of the 800 census tracts in Chicago. The remaining 150 had to be manually matched. Once each tract was matched to the appropriate Community Area I then consolidated the ACS statistics for each Community Area. Once this was completed the appropriate calculations on each predictor were finally made. The ACS shapefile was joined to the shapefile with crime and the other six predictors and there was finally a master shapefile with all of the outcome and predictor variables. Now that all of the data was collected and in an ideal form the next steps were to calculate descriptive statistics and to run regression models. Table 4 displays the programs that were used throughout the study.

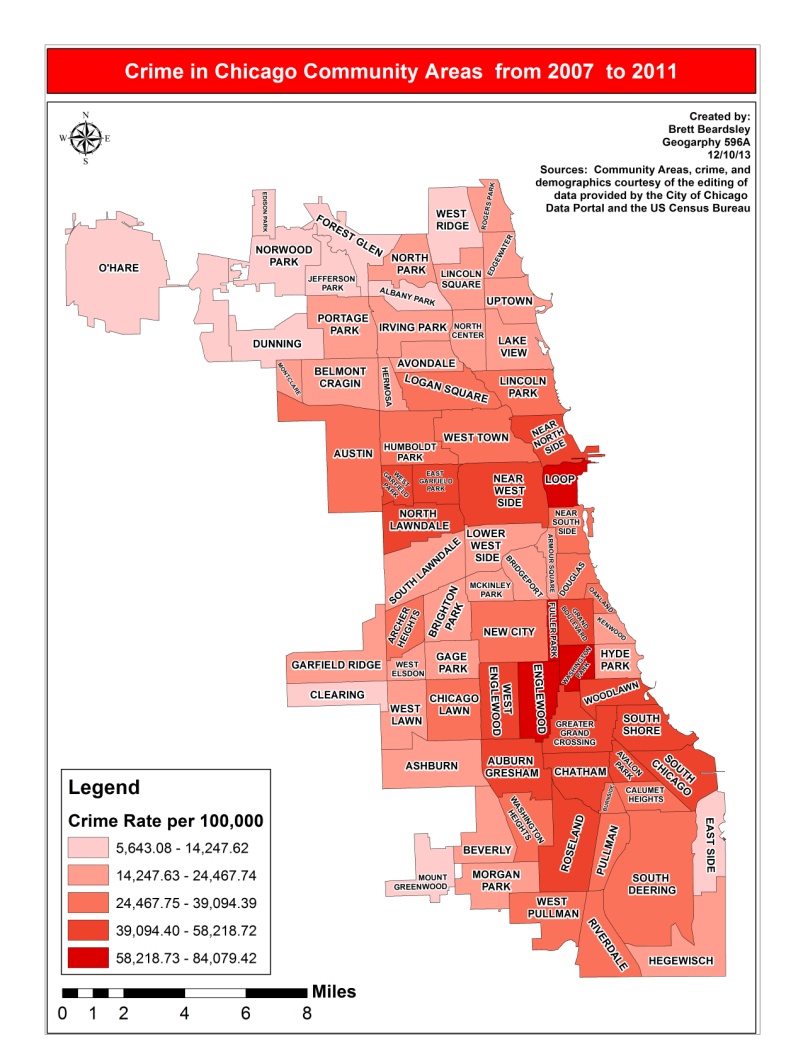
**Table 4: Programs used in study**

|  |  |
| --- | --- |
| Program | Tasks |
| Excel | Data collection, exploration, and cleaning |
| ArcGIS | Data manipulation, choropleth maps, and some descriptive statistics |
| Minitab | Correlation matrices |
| GeoDa | Spatial Weights Matrix, Moran's I, OLS, spatial lag, and error regression, and residual maps and scatter plots |

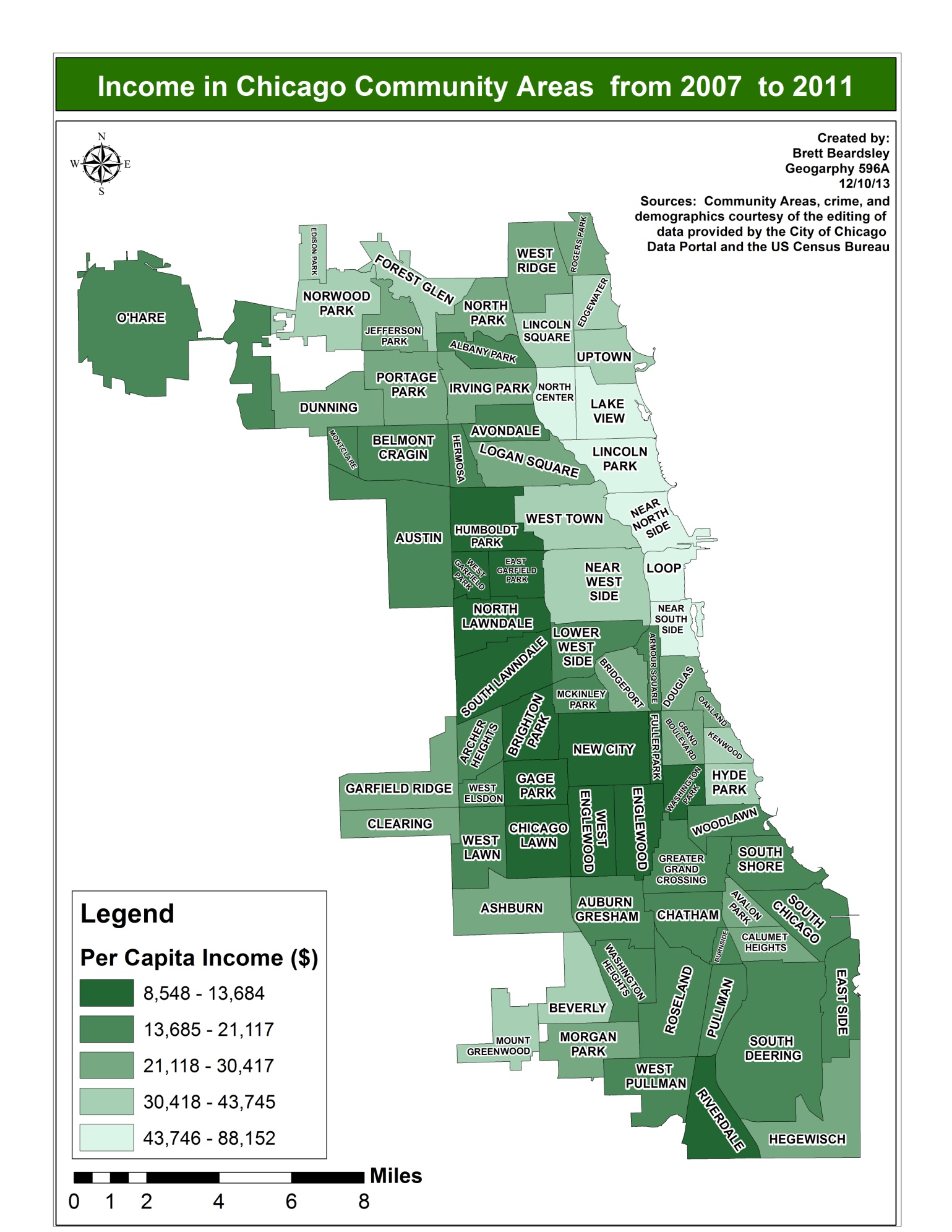
# 

# Data Analysis

The next step was to examine the data and create basic maps of what the gathered data would display to get a better understanding of it and to see any patterns. Figure 2 shows a choropleth map of total crime rate in each Community Area. The west and south sides of Chicago are the areas that are notoriously known as having high crime and Figure 2 supports this. The north side is generally known as the safest part of Chicago and this is supported by Figure 2 as well. One interesting finding from the map is that the Loop has a very high crime rate. The Loop is the central business district of Chicago. When the high rate of crime was discovered the first thought was that it has be property crime related because there is a very large number of wealthy people in the Loop every day and if violence were a problem in the Loop there would be a very large amount of publicity about it because it is the heart of Chicago. The people that actually live in the Loop also are likely to have a good amount of money because property there is expensive. Figure 3 shows per capita income by Community Area. It supports the theory that there is a large amount of money in the Loop as you can see that the Loop is in the top 5 for highest per capita income in the city. Property crime cannot occur if there is no property worth taking. When examining property crime rates by Community Area the Loop, as was hypothesized, had far and away the highest property crime rate at 15,553 property crimes per 100,000 people. The next highest was over 3,000 less at 12,108 in Fuller Park. See Appendix A for a selection of choropleth maps for other variables in the study.



**Figure 2: Total Crime rate in Chicago Community Areas**

**Figure 3: Per Capita income in Chicago Community Areas**

## Descriptive Statistics

# 

# The descriptive statistics give a view of the Community Areas that have the highest and lowest values for crimes and predictor variables. They also display the means of all of the variables, the variation across Chicago of each variable, and the autocorrelation of each variable. The statistics were calculated in ArcMap and GeoDa. Table 5a and 5b display the descriptive statistics for the dataset. In terms of crime you can see that Edison Park, O'Hare, and Mount Greenwood were the only Community Areas that had a minimum rate for any crime, with six minimum crime rates each for O'Hare (murder and non negligent manslaughter, aggravated assault, rape, burglary, arson, and violent crime) and Edison Park (murder and non negligent manslaughter, robbery, motor vehicle theft, larceny/theft, total crime, and property crime) and one for Mount Greenwood (murder and non negligent manslaughter). These neighborhoods are all known as low crime areas, so it was not surprising to see them have some minimum crime rate values. Mount Greenwood and Edison Park also appear on the minimum and maximum values for the predictors, which can give us some insight into what predictors are associated with low crime and alert us to trends to examine when conducting more in depth analysis. Based on minimum values for African Americans in Edison Park, households below poverty in Edison Park and Mount Greenwood, and housing units renter occupied in Mount Greenwood it appears that low levels of murder and non negligent manslaughter, robbery, motor vehicle theft, larceny/theft, total crime, and property crime could be related to low levels of those predictor variables. Edison Park also has the highest percentage of whites, so that could be associated with low levels of murder and non negligent manslaughter, robbery, motor vehicle theft, larceny/theft, total crime, and property crime as well without any further analysis performed. It is also important to note that Edison Park has the lowest total crime rate of any Chicago Community Area from 2007 to 2011.

When examining the Community Areas with the maximum values for crimes you find that West Garfield Park, Washington Park, Englewood, Fuller Park, Loop, and Burnside have maximum rates, with four for Fuller Park (robbery, motor vehicle theft, total crime, and violent crime), two each for Englewood (rape and burglary) and the Loop (larceny/theft and property crime), and finally one each for West Garfield Park (murder and non negligent manslaughter), Washington Park (aggravated assault), and Burnside (ironically arson). These were more surprising than the low crime areas because only West Garfield Park and Englewood are widely known as being among the worst of the worst areas in the city. As was previously mentioned the Loop was startling at first, but after examining it in greater depth it really was not that surprising. Burnside, Loop, and West Garfield Park also have some of theminimum and maximum values for the predictors, which can give some ideas of possible relationships between specific types of crime and crime in general and the predictors. Based on minimum values for foreign born in West Garfield Park, whites in Burnside, and three or more bedrooms in housing units and population aged under 18 or over 64 in the Loop it appears that high levels of murder and non negligent manslaughter, larceny/theft, property crime, and arson could be related to low levels of those predictor variables. Not surprisingly Burnside also had the maximum value for African Americans, which shows a possible relationship between high African American populations and high levels of arson. It is also important to note that Fuller Park had the highest total crime rate of any Chicago Community Area from 2007 to 2011. This was surprising because in my 30 years in the Chicagoland area I had never heard of it before this study.

Studying which Community Areas have minimum and maximum values for the predictor variables also sheds some light on some possible trends in the data set. The community area that had the most combined minimum and maximum values was Riverdale with seven (minimum for per capita income and median value of owner occupied housing units and maximum for female-headed households, household below poverty, vacant housing units, renter occupied housing units, and percent of population under 18 or over 64) and they were almost all predictors that would presumably be associated with high crime rates.The lone exception was having the most people under 18 or over 64. The fact that Riverdale did not have any corresponding maximumvalues for crime is surprising with the level of poverty and low income there. Even more surprising is that Riverdale was ranked 25th in total crime rate. A further analysis of how Riverdale managed to not have any maximum crime rates and be 25th in total crime with seven predictors being this high or low would be interesting and could lead to ideas of how other impoverished Community Areas could lower crime rates. A table showing the total crime rate and ranking of each Community Area can be found in Appendix B.

The Moran's I values will indicate if there's spatial autocorrelation in the variables. Spatial autocorrelation is the fact that data from locations near one another in space are more likely to be similar than data from locations remote from one another. The value ranges from -1 to +1. In general a Moran's I of 0.3 or more or of -+0.3 or less indicates a relatively strong autocorrelation (Sullivan & Unwin, 2010). Please note that a first order Queen's spatial weights matrix was created before the Moran's I values were calculated. It was not surprising then to see that every variable in the study other than population aged 15 to 24 had strong positive spatial autocorrelation. It is important to note that each Moran's I value was found to be significant. The strong positive spatial autocorrelation for all of the variables shows that the variables tended to be similar to one another in nearby geographic locations. It is also important to note that a spatial weights matrix was created prior to calculating the Moran's I values. The Coefficient of Variation (CV) was also included in the descriptive statistics. It is calculated by dividing the standard deviation by the mean for each variable. It is a measure of a variable's variability across an area. As you can see the highest CVs in terms of crimes are for the violent crimes. This means that there is a larger amount of variation among violent crimes throughout Chicago from Community Area to Community Area than there is for property crimes. This is intuitive as property crime is much more common throughout any area than violent crime. Many more

**Table 5a: Descriptive Statistics for Outcome Variables**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Minimum | Maximum | Mean | Std. Deviation | Moran's I | Coefficient of Variation |
| Murder and Non Negligent Manslaughter | 0-Mt. Greenwood, O'Hare, and Evergreen Park | 78.04-West Garfield Park | 19.34 | 20.46 | 0.529 | 1.06 |
| Aggravated Assault | 27.65-O'Hare | 2,193.51-Washington Park | 637.33 | 582.08 | 0.583 | 0.91 |
| Forcible Rape | 5.43-O'Hare | 164.98-Englwood | 53.08 | 41.27 | 0.58 | 0.78 |
| Robbery | 17.8-Edison Park | 2,454.87-Fuller Park | 601.69 | 522.23 | 0.573 | 0.87 |
| Burglary | 5.78-O'Hare | 2,832.23-Englewood | 1,005.04 | 613.58 | 0.648 | 0.61 |
| Motor Vehicle Theft | 78.33-Edison Park | 1,971.12-Fuller Park | 702.28 | 422.23 | 0.587 | 0.60 |
| Larceny/Theft | 765.47-Edison Park | 14,719.27-Loop | 2,917.09 | 1,970.91 | 0.45 | 0.68 |
| Arson | 0.99-O'Hare | 71.13-Burnside | 23.17 | 17.26 | 0.536 | 0.74 |
| Total Crime | 1,128.62-Edison Park | 16,815.88-Fuller Park | 5,959.02 | 3,456.51 | 0.553 | 0.58 |
| Violent Crime | 52.83-O'Hare | 4,707.58-Fuller Park | 1,311.44 | 1,145.19 | 0.593 | 0.87 |
| Property Crime | 1,053.85 Edison Park | 15,552.81-Loop | 4,647.59 | 2,515.41 | 0.515 | 0.54 |

**Table 5b: Descriptive Statistics for Predictor Variables**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Variable | Minimum | Maximum | Mean | Std. Deviation | Moran's I | Coefficient of Variation |
| African American | 0.24%-Edison Park | 99.5%-Burnside | 40.27% | 39.95 | 0.703 | 0.99 |
| Households Built After 1989 | 0.87%-Pullman | 27.7%-Near South Side | 5.77% | 5.19 | 0.563 | 0.90 |
| Foreign Born | 0.26%-West Garfield Park | 53.9%-Armour Square | 19.21% | 15.31 | 0.586 | 0.80 |
| Overcrowded Households | 0.3%-North Center | 17.2%-North Lawndale | 4.94% | 3.67 | 0.462 | 0.74 |
| White | 0.5%-Burnside | 97.33%-Edison Park | 41.84% | 30.95 | 0.675 | 0.74 |
| Per Capita Income | $8,548-Riverdale | $88,152-Near North Side | $25,586.65 | 15,096.60 | 0.583 | 0.59 |
| Female-Headed Households | 2.6%-Lake View | 52.45%-Riverdale | 21.06% | 12.35 | 0.601 | 0.59 |
| 25+ without a High School Diploma | 2.8%-Lake View and Near North Side | 54.8%-North Lawndale | 20.77% | 11.87 | 0.541 | 0.57 |
| Households Below Poverty | 3.5%-Edison Park and Mount Greenwood | 58.4%-Riverdale | 20.84% | 11.54 | 0.398 | 0.55 |
| Population Density | 1,529.72-South Deering | 31,768.39-Edgewater | 13,000.08 | 6,574.63 | 0.511 | 0.51 |
| 16+ Unemployed | 4.96%-North Center | 35.9%-Englewood | 14.33% | 7.15 | 0.415 | 0.50 |
| Vacant Housing Units | 4.53%-Ashburn | 36.24%-Riverdale | 14.1% | 6.94 | 0.46 | 0.49 |
| Median Value of Owner Occupied Housing Units | $106,050-Riverdale | $552,920-Lincoln Park | $253,307.64 | 98,036.17 | 0.743 | 0.39 |
| Housing Units Renter Occupied | 10.53%-Mount Greenwood | 84.43%-Riverdale | 50% | 18.57 | 0.532 | 0.37 |
| 3+ Bedrooms Housing Units | 8.49%-Loop | 81%-Ashburn | 45.8% | 15.83 | 0.606 | 0.35 |
| Population 15-24 | 7.03%-Near South Side | 22.48%-Hyde Park | 14.74% | 3.12 | 0.12 | 0.21 |
| Aged Under 18 or Over 64 | 14.2%-Loop | 51%-Riverdale | 35.79% | 7.23 | 0.64 | 0.20 |

people are willing to steal from people when they are not present than hurt them. It is a much less risky and more compassionate, for lack of a better term, type of crime than physically hurting or threatening another person. Once an analysis of the descriptive statistics was complete it was time to create correlation matrices of all of the variables.

## Correlation Matrices

The next step in the study was to create correlation matrices in GeoDa among the outcome variables, the predictor variables, and between the variables. The purpose of this was to test the strength and significance of the relationship between variables. If it is found that two variables are strong and significantly correlated with one another than one should be removed from the analysis because it is essentially like testing the same thing twice and will also introduce multicollinearity into my regression models. This is critical because multicollinearity causes major problems in regression models and needs to be eliminated as much as possible. The correlation coefficient is a number between -1 and +1 the measures the degree of which two variables are related. The farther the number is away from one, the stronger the relationship is between two variables (Stockburger, 1998). The threshold used in this study to indicate if a relationship was too strong between two variables was positive or negative 0.75.

To review, at the beginning of this study there were 11 outcome variables and 17 predictor variables. Given that crime usually is clustered in areas it was expected that there would be correlation amongst the outcome variables and the number of variables would decrease significantly. The first matrix created was among the specific crime variables. Violent, property, and total crime were not included because there was a goal of trying to use specific crime types. There also would obviously be strong correlation between the crimes that the total crime rates consisted of, so there was no reason to put the individual crimes with the total crime rates. The strong and significant correlation amongst the outcome variables was higher than expected. Aggravated assault, rape, robbery, burglary, motor vehicle theft, and arson were all significantly correlated with at least one other outcome variable at a level of positive 0.75 in every case, but robbery and arson. In fact the lowest any of the correlation coefficients were was very high at 0.823. This high level of correlation between so many of the variables would have left the model with only arson and larceny as outcome variables and no violent crimes. The complete correlation matrix among individual crime types can be found in Appendix B. From this high level of multicollinearity and the lack of possible variables to use it was decided that violent crime, property crime, and total crime would be the outcome variables for the model and the other eight predictor variables would not be used.

The correlation matrix for the predictors was created next. The same threshold of positive or negative 0.75 was used to eliminate predictors that were too highly correlated with one another. Percentage below poverty, percent female headed households, and percent black were all eliminated because of significant and higher correlations with other predictors. Mean household value, 16 and over and unemployed, and 25 and over and no high school diploma also had strong and significant correlations with certain predictor variables, but were kept because of the value and uniqueness of what they measured. Mean household value was the only variable concerning home worth, 16 and over and unemployed was the only employment measures, and 25 and over and no high school diploma was the only education measure. It was also strongly believed that these variables would have a relationship with crime and should not have been removed. The complete correlation matrix among individual crime types can be found in Appendix B. Next the remaining thirteen predictor variables were put into a correlation matrix with the three outcome variables. Percent of overcrowded households, population density, percent of households with three or more bedrooms, percent of households built after 1989, and percent population under 18 and over 64 were eliminated because of a lack of a significant and moderate correlation with the property, violent, or total crime. This now left three outcome variables, violent crime rate, property crime rate, and total crime rate, and nine predictor variables, percent aged 16 and over and unemployed, percent 25 and over without a high school diploma, per capita income, percent white, percent aged 15 to 24, percent vacant housing units, percent foreign born, percent housing units renter occupied, and mean household value.

At this point a correlation matrix with the three outcome variables was created to confirm that the expected multicollinearity was present. That matrix along with a consolidated version of the matrix among remaining predictors and a matrix with the remaining outcome and predictor variables can be found in Tables 6, 7 and 8. As you can see there is a significant, strong correlation among the outcome variables. As a result of this it was then decided that there would be three regression models, one each for total crime, violent crime, and property crime. This is because creating one model with all three outcome variables would contain multicollinearity, which causes problems in regression models. Now that the predictor and outcome variables have been narrowed down it is time to see which predictors can best be used to predict the violent, property, and total crime rates in Chicago Community Areas.

**Table 6 : Correlation Matrix of Selected Outcome Variables**

|  |  |  |
| --- | --- | --- |
| Outcome Variable | Violent Crime | Property Crime |
| Property Crime | 0.748 |  |
| p-value | 0.000 |  |
|  |  |  |
| Total Crime | 0.876 | 0.976 |
| p-value | 0.000 | 0.000 |

**Table 7 : Correlation Matrix of Selected Predictor Variables**

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Variables | 16 and Over and Unemployed | 25 and Over and No High School Diploma | Per Capita Income | White | 15 to 24 | Vacant | Foreign Born | Renters |
| 25 and Over No High School Diploma | 0.367 |  |  |  |  |  |  |  |
| p-value | 0.001 |  |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| Per Capita Income | -0.643 | -0.709 |  |  |  |  |  |  |
| p-value | 0.000 | 0.000 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| White | -0.705 | -0.226 | 0.557 |  |  |  |  |  |
| p-value | 0.000 | 0.048 | 0.000 |  |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| 15 to 24 | 0.388 | 0.457 | -0.385 | -0.361 |  |  |  |  |
| p-value | 0.000 | 0.000 | 0.001 | 0.001 |  |  |  |  |
|  |  |  |  |  |  |  |  |  |
| Vacant | 0.652 | 0.260 | -0.312 | -0.604 | 0.522 |  |  |  |
| p-value | 0.000 | 0.023 | 0.006 | 0.000 | 0.000 |  |  |  |
|  |  |  |  |  |  |  |  |  |
| Foreign Born | -0.363 | 0.500 | -0.077 | 0.508 | -0.007 | -0.436 |  |  |
| p-value | 0.001 | 0.000 | 0.508 | 0.000 | 0.949 | 0.000 |  |  |
|  |  |  |  |  |  |  |  |  |
| Renters | 0.330 | 0.192 | -0.146 | -0.460 | 0.440 | 0.698 | -0.110 |  |
| p-value | 0.003 | 0.095 | 0.206 | 0.000 | 0.000 | 0.000 | 0.343 |  |
|  |  |  |  |  |  |  |  |  |
| Mean Household Value | -0.743 | -0.460 | 0.811 | 0.701 | -0.363 | -0.417 | 0.235 | -0.062 |
| p-value | 0.000 | 0.000 | 0.000 | 0.000 | 0.001 | 0.000 | 0.040 | 0.595 |

**Table 8 : Correlation Matrix of Selected Outcome and Predictor Variables**

|  |  |  |  |
| --- | --- | --- | --- |
| Variable | Total Crime | Violent Crime | Property Crime |
| 16 and Over Unemployed | 0.544 | 0.756 | 0.404 |
| p-value | 0.000 | 0.000 | 0.000 |
|  |  |  |  |
| 25 and Over No High School Diploma | -0.005 | 0.186 | -0.091 |
| p-value | 0.967 | 0.106 | 0.431 |
|  |  |  |  |
| Per Capita Income | -0.097 | -0.437 | 0.066 |
| p-value | 0.400 | 0.000 | 0.571 |
|  |  |  |  |
| White | -0.654 | -0.816 | -0.527 |
| p-value | 0.000 | 0.000 | 0.000 |
|  |  |  |  |
| 15 to 24 | 0.388 | 0.440 | 0.332 |
| p-value | 0.000 | 0.000 | 0.003 |
|  |  |  |  |
| Vacant | 0.747 | 0.836 | 0.647 |
| p-value | 0.000 | 0.000 | 0.000 |
|  |  |  |  |
| Foreign Born | -0.589 | -0.622 | -0.526 |
| p-value | 0.000 | 0.000 | 0.000 |
|  |  |  |  |
| Renters | 0.502 | 0.533 | 0.447 |
| p-value | 0.000 | 0.000 | 0.000 |
|  |  |  |  |
| Mean Household Value | -0.344 | -0.579 | -0.210 |
| p-value | 0.002 | 0.000 | 0.067 |

## Regression

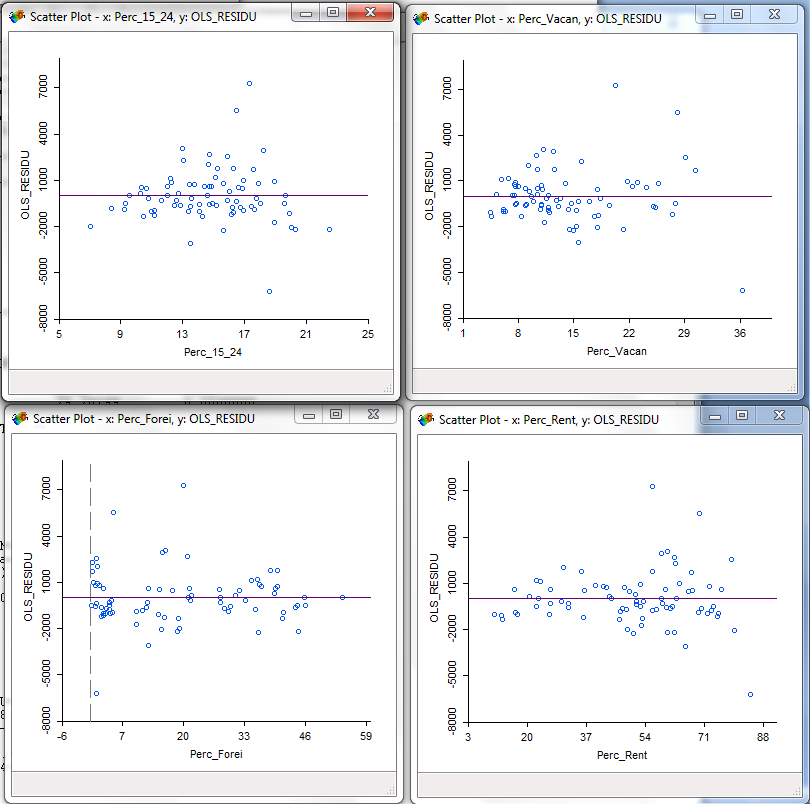
Regression analysis is a statistical tool that seeks to find the relationship between dependent variables and one or more independent variables (Sykes, 1992). It essentially will calculate what percentage of the variation in an outcome variable can be predicted by a set of predictor variables. A very useful characteristic of regression modeling is that it will also indicate the strength of the relationship between the dependent variable and each independent variable, while controlling for the other independent variables (Adams, 2007). For example in this study one can see the relationship between total crime and percentage of foreign born people in a Chicago Community Area while controlling for the other eight predictor variables. Regression modeling will therefore help reach the study's goals of identifying which predictors are most associated with violent, property, and total crime. This will also help identify any patterns among predictors between the different crime types. Lastly the regression modeling will also help identify, which predictors have the strongest influence each crime type. Please note that all of the Ordinary least Squares (OLS) regression models discussed used a first order Queen's spatial weights matrix.

### OLS Regression-Total Crime Rate

Ordinary Least Squares (OLS) regression is the proper starting point for all spatial regression analysis. It creates a global model of the variable one is trying to understand or predict (ESRI, n.d.). The outcome variables that are trying to be understood in this study are total crime, violent crime, and property crime in Chicago Community Areas. It was also decided that each of those outcome variables would have their own regression model. The first one that was run was the OLS Regression model for total crime.

The first OLS Regression model that was run for total crime had all remaining nine predictor variables, percent of the population aged 15 to 24, percent 16 and over and unemployed, percent 25 and over and no high school diploma, per capita income, percent white, percent vacant housing units, percent foreign born, percent housing units occupied by renters, and mean household value. An OLS regression model has many diagnostics for a variety of potential problems. Among these diagnostics is one for mutlicollinearity. Minimizing multicollinearity is crucial to any OLS regression model, so it is vital that it is below a certain threshold. A multicollinearity condition number over 30 indicates too much multicollinearity and this first model was just over 30. This was not surprising because as was mentioned previously there were still predictor variables that were highly correlated with one another. The residuals that the model produced for total crime were also way too large. This indicates that the model was not a good one because residuals are the differences between actual values for total crime in Chicago Community areas and the values that the model predicted. The smaller the residuals the better the model. The multicollinearity and large residuals meant at least one predictor variable needed to be removed. Mean household value was removed because it had strong correlation with three other predictor variables and it was not significant in predicting the total crime rate.

The second model for total crime had the same predictor variables as the first minus mean household income and this time the multicollinearity number was 27.99, so the multicollinaearity was at a good level. The Lagrange Multiplier and Robust LM for lag were both significant in this model. These two diagnostics indicate spatial dependence for the dependent variable and call for a spatial lag regression model to be run. The residuals in the lag regression model were still too large, so more predictors had to be removed. Percentage of foreign born and percentage of renters were removed after this model partly because neither were significant predictors of total crime and they also were strongly correlated with other predictor variables. They also displayed heteroscedasticity, which means that they had a non-constant error variance. This is a problem because it can corrupt the results of the model. Heteroscedasticity can be detected by examining scatter plots of the residual values against values of each variable. If the residuals go up or down as the variable's values increase or decrease then the variable has heteroscedasticity. Figure 4 shows the scatter plot for percentage of renters and the residuals. You can see that as the percentage of renters increases, so too do the residuals.

**Figure 4: Scatter plot of percent renters and residuals**

The third regression model for total crime had six predictor variables. The multicollinearity decreased to 22.48 in this model. Once again a spatial lag regression model was called based on the diagnostics. There were still three predictors that were not significant predictors of total crime. When the variables are not significantly contributing to predicting the outcome variable that means that they are not adding much value to the model. Percent of the population over the age of 16 that is unemployed was one of the non-significant variables and it was removed. The fourth regression model included five predictor variables, percentage of the

population aged 15 to 24, percent aged 25 and over without a high school diploma, per capita income, percent white, and percent vacant housing units. The fourth model's multicollinearity condition was 18.68 which was 18.68 which was an improvement. It called for a spatial lag regression as well and still had two predictor variables that were not significant in predicting total crime. Once again this called for the removal of at least one of them and this time percent of the population aged 25 and over without a high school diploma was removed. The fifth regression model now included four variables, percentage of the population aged 15 to 24, per capita income, percent white, and percent vacant housing units. The multicollinearity condition number went down again to 17.17. Every predictor variables but percentage 15 to 24 was significant. Percentage of the population aged 15 to 24 was kept because it was the only variable that did not have strong autocorrelation. It also is a good variable to keep because crime is often committed by people in that age group. A satisfactory regression model for total crime had finally been reached after iterating through five models. The final model had the following predictor variables: percentage of the population aged 15 to 24, per capita income, percentage white, and percentage of housing units that are vacant.

Now that there is a final model for total crime rate I will go into a more specific analysis on more of the diagnostics from the regression model. Table 9 will illustrate many of the numbers that are going to be discussed.

**Table 9: Total Crime OLS Regression Results**

|  |  |  |  |
| --- | --- | --- | --- |
| Diagnostic | Coefficient | Value/T-statistic | Probability |
| Constant | 1,330.20 | 0.96 | 0.205 |
| Per Capita Income | 0.09 | 5.00 | 0.000 |
| Percent White | -60.23 | -6.01 | 0.000 |
| Percent 15 to 24 | 91.00 | 1.08 | 0.285 |
| Percent Vacant | 249.58 | 5.78 | 0.000 |
| Multicollinearity Condition Number |  | 17.17 |  |
| Log likelihood |  | -687.55 |  |
| Akaike info criterion |  | 1,385.09 |  |
| Schwarz criterion |  | 1,396.81 |  |
| R-squared |  | 0.72 |  |
| F-statistic |  | 46.45 | 3.07e-19 |
| Moran’s I |  | 1.53 | 0.126 |
| Lagrange Multiplier (lag) |  | 4.10 | 0.043 |
| Robust LM (lag) |  | 3.75 | 0.053 |
| Lagrange Multiplier (error) |  | 0.94 | 0.333 |
| Robust LM (error) |  | 0.58 | 0.445 |

There are numerous other diagnostics found in a regression model that test various characteristics. One is for the model's goodness of fit. One of the diagnostics for that is the r-squared value, which is a goodness of fit test. The r-squared for the regression model for total crime rate was 0.72. This means that 72% of the variation in the total crime rate in Chicago Community Areas was predicted or explained by the predictor variables in the model. That is a high percentage and shows that my model is a good one for predicting total crime rate in a given Chicago Community Area. The F-statistic tests the null hypothesis that all regression coefficients are jointly zero. The model's F-statistic of 46.45 and it's probability of 3.08e-19 reject this null hypothesis. The log likelihood, Akaike info criterion, and Schwarz criterion are used to maintain comparability with the fit of the spatial error and spatial lag regression models if they are run. The closer the values are to zero, the better the fit of the model.

The coefficients of each variable show what direction and how much of an affect each predictor has on the total crime rate if all of the other predictors are held constant. For example the coefficient for percent housing units vacant is 250. This means that for every percent increase in vacant housing units the total crime rate increases by 250 crimes per 100,000 people. From Table 9 you can see that as per capita income, percentage of the population between 15 and 24, and percent vacant housing units increased so did the crime right. You can also see that as the percentage of white people increased the total crime rate decreased. It is also important to look at the t-statistic and its significance to see if these values mean anything. The T-statistic is the coefficient divided by the standard error and the probability measures its significance to the model. From the table you can see that per capita income, percent white, and percent vacant housing units are significant. This confirms that increases in per capita income and percent vacant housing units increases the total crime rate and that an increase in percent white decreases total crime rate. You can also see that vacancy has the most significant affect on total crime rate of the predictors that used percentages. You can also see that percent of the population aged 15 to 24 is not significant to the model, but it is being kept in the model because of its lack of autocorrelation and its theoretical importance.

The next section is the diagnostics for multicollinearity and spatial dependence. As was previously mentioned a multicollinearity condition number above 30 indicates too much mulitcollinearity, so this model's number of 17.17 means multicollinearity is not a problem. The Moran's I value measures autocorrelation and for this model it is not significant with a probability of 0.13, which means autocorrelation is not a problem. The Lagrange Multiplier test statistics help the user specify whether or not an alternative specification for their regression model is needed (Anselin, 2005). You first need to look at the Lagrange Multiplier lag value. If and only if it is significant you can then move onto the Robust LM (lag) number. If the Robust LM lag value is also significant then a lag model should be run. The same procedure should be followed with the Lagrange Multiplier error and LM error values. In the total crime model the Lagrange Multiplier lag value and Robust LM model are significant with probabilities of 0.04 and 0.05 respectively. This means that a spatial lag model is needed. You should also check on the Lagrange Multiplier error value just in case it is also shown to be significant. For the total crime model it is 0.33, which is not significant.

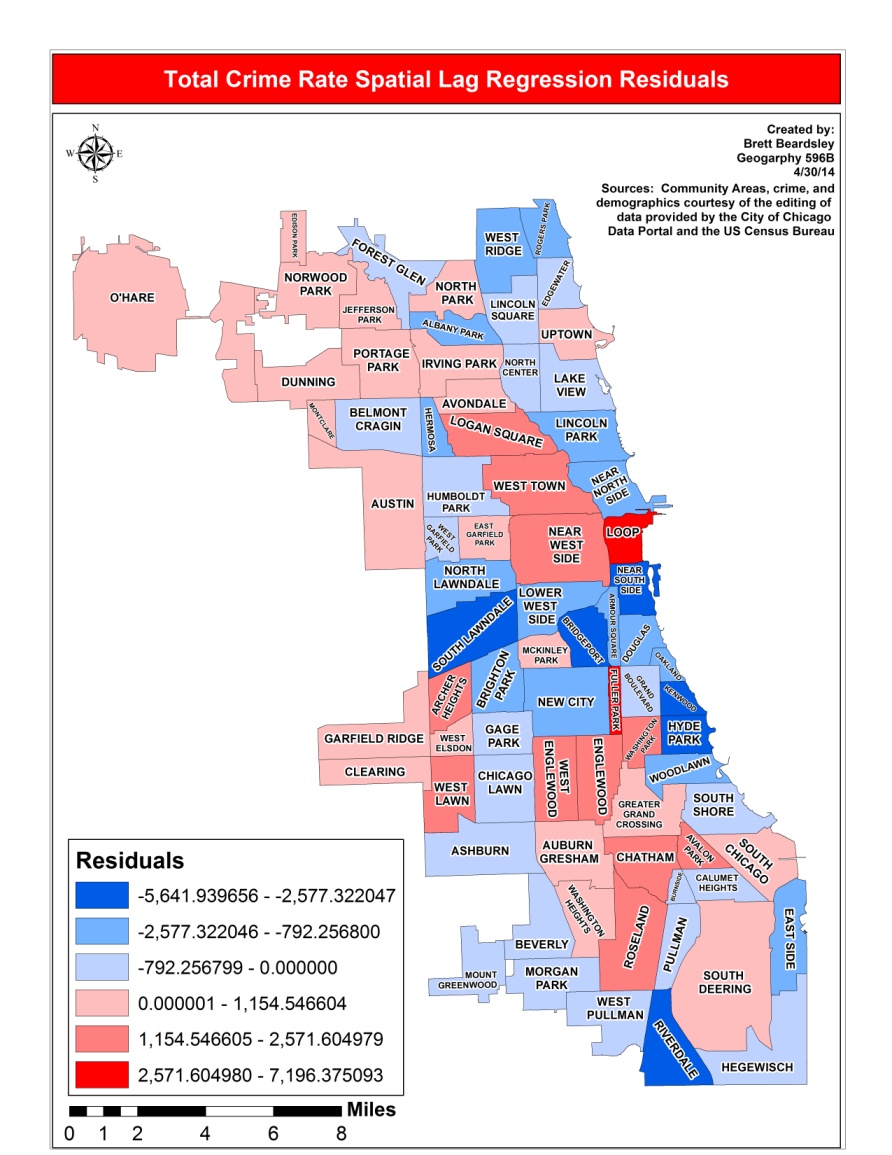
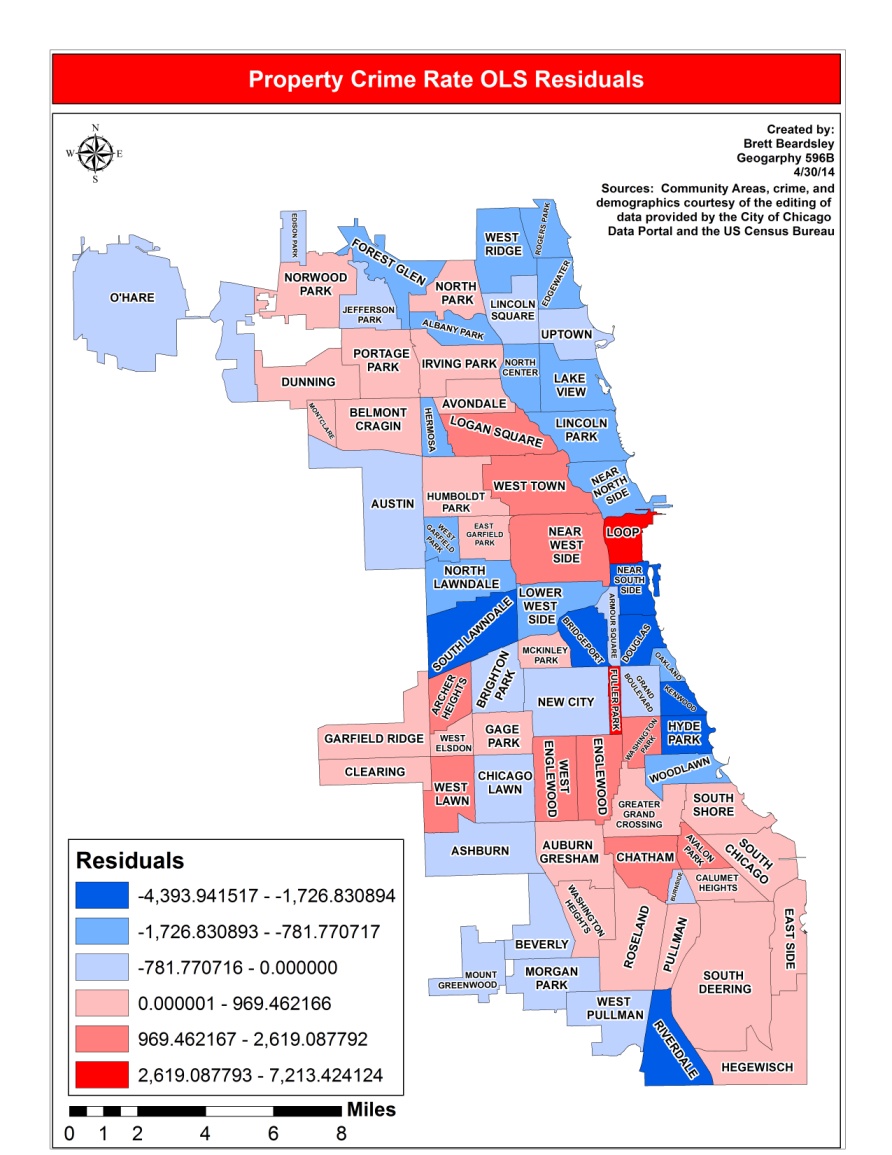
### Spatial Lag Model-Total Crime Rate

A spatial lag model assumes that there is spatial dependence among the outcome variable values. In this case that means that the model will be assuming that the total crime rate in one Community Area will be affected or related to the total crime in the surrounding Community Areas. In order to account for this spatial dependence a spatial lag model adds a spatial lag coefficient to the model. You can see the spatial lag coefficient of 0.31 in Table 10.

**Table 10: Total Crime Spatial Lag Regression Results**

|  |  |  |  |
| --- | --- | --- | --- |
| Diagnostic | Coefficient | Value/Z-test | Probability |
| Constant | 231.36 | 0.17 | 0.864 |
| Per Capita Income | 0.07 | 3.62 | 0.000 |
| Percent White | -42.85 | -3.80 | 0.000 |
| Percent 15 to 24 | 64.71 | 0.82 | 0.415 |
| Percent Vacant | 215.45 | 5.24 | 0.000 |
| Spatial lag term | 0.31 | 2.73 | 0.006 |
| Log likelihood |  | -685.01 |  |
| Akaike info criterion |  | 1,382.01 |  |
| Schwarz criterion |  | 1,396.08 |  |

The coefficient of each predictor variable in the spatial lag model decreased from the OLS regression model. You can also see that vacancy also had the largest affect on total crime rate with a coefficient of 215.45. The direction and significance of each is also the same as in the OLS regression. model. The values for the log likelihood, Akaike info criterion, and Schwarz criterion all became closer to zero from the OLS regression model, which indicates a better goodness of fit for the spatial lag regression model. Taking a look at the residual maps in Figure 5 between the two models also can help one visualize the better fit of the model. You can see that there are less extreme low residuals and the same number of extreme high residuals in the spatial lag residual map. There are also less values in the second highest and lowest residual categories in the spatial lag residual map compared to the OLS residual map. This means the spatial lag model more accurately predicts the total crime rate in Chicago Community Areas than the OLS model. The next model that was run was the OLS model for violent crime.

 **Figure 5: OLS and Spatial Lag residuals**

### OLS Regression-Violent Crime Rate

The first OLS regression model for violent crime had all nine predictor variables that made it through the correlation matrix reductions. This model had too much multicollinearity at 30.11, so second model had to be run. Mean household value was removed because it was insignificant and had strong and significant correlations with some of the remaining predictor variables. The multicollinearity in the second model was at a good level at 27.99, but there were still some variables that were insignificant to the violent crime rate. Twenty five and over without a high school diploma, unemployed, and renters were removed because of non-significance and heteroscedasticity. This left five variables for the third model: per capita income, percentage of the population that is white, percentage of the population aged 15 to 24, percentage of housing units that are vacant, and percentage of the population that is foreign born. The multicollinearity condition number went down again to 18.45 and every predictor other than population aged 15 to 24 was significant. Percentage of the population aged 15 to 24 was kept because it was the only variable that did not have strong autocorrelation. It also is a good variable to keep because crime is often committed by people in that age group. A satisfactory regression model for violent crime had been reached after iterating through three models. The final model for violent crime rate had the following variables: per capita income, percentage of the population that is white, percentage of the population aged 15 to 24, percentage of housing units that are vacant, and percentage of the population that is foreign born.

Now that there is a final model for violent crime rate I will go into a more specific analysis on more of the diagnostics from the regression model. Table 11 will illustrate many of the numbers that are going to be discussed. The r-squared value of 0.89 indicates that 89% of the variation in the violent crime rate in Chicago Community Areas was predicted or explained by the predictor variables in the model. That is a very high percentage and shows that my model is a great one for predicting violent crime in a given Chicago Community Area. The F-statistic of 110.71 with a probability of 4.22e-32 rejects the null hypothesis that all regression coefficients are jointly zero. Both of these diagnostics show a good, goodness of fit for my model.

**Table 11: Violent Crime OLS Regression Results**

|  |  |  |  |
| --- | --- | --- | --- |
| Diagnostic | Coefficient | Value/T-statistic | Probability |
| Constant | 1,153.85 | 3.85 | 0.000 |
| Per Capita Income | -0.01 | -2.17 | 0.033 |
| Percent White | -11.67 | -4.59 | 0.000 |
| Percent 15 to 24 | 10.07 | 0.54 | 0.588 |
| Percent Vacant | 78.95 | 8.26 | 0.000 |
| Percent Foreign | -19.68 | -4.68 | 0.000 |
| Multicollinearity Condition Number |  | 18.45 |  |
| Log likelihood |  | -567.88 |  |
| Akaike info criterion |  | 1,147.77 |  |
| Schwarz criterion |  | 1,161.83 |  |
| R-squared |  | 0.89 |  |
| F-statistic |  | 110.71 | 4.22e-32 |
| Moran’s I |  | 4.29 | 0.000 |
| Lagrange Multiplier (lag) |  | 2.29 | 0.130 |
| Robust LM (lag) |  | 0.14 | 0.712 |
| Lagrange Multiplier (error) |  | 11.67 | 0.002 |
| Robust LM (error) |  | 9.51 | 0.002 |

Percent vacant housing units had the largest coefficient at 78.95, which means that for every one percent vacant housing increases the violent crime rate will increase by 78.95 crimes per 100,000 people. This also means that it has a larger affect on violent crime than percent white, percent 15 to 24, and percent foreign born. As you can see in table 11 as the percentage of people aged 15 to 24 and the percentage of vacant housing units increase, so does violent crime. Therefore these variables have a positive relationship with violent crime. The negative coefficients of per capita income, percentage white, and percent foreign born indicate that as each of those predictors increase violent crime decreases. Therefore these variables have a negative relationship with violent crime. The probabilities of the T-statistics indicate that all predictors other than percent aged 15 to 24 are significant. As was previously mentioned the multicollinearity number of 18.45 means that multicollinearity is not a problem. The significant Moran's I value means that autocorrelation is a problem. The Lagrange Multiplier value for lag is not significant at 0.13, which means the model does not need a spatial lag regression specification. The Lagrange Multiplier and Robust LM values for error are significant at of 0.002 and 0.002. This calls for a spatial error regression model specification to be run.

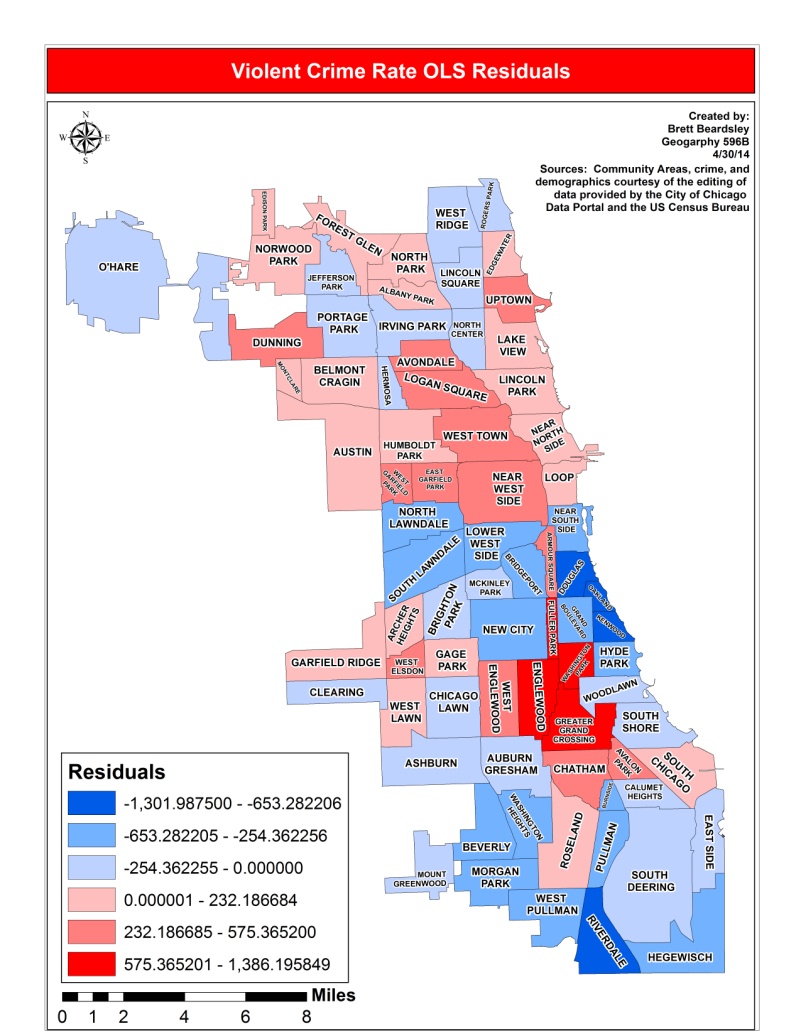
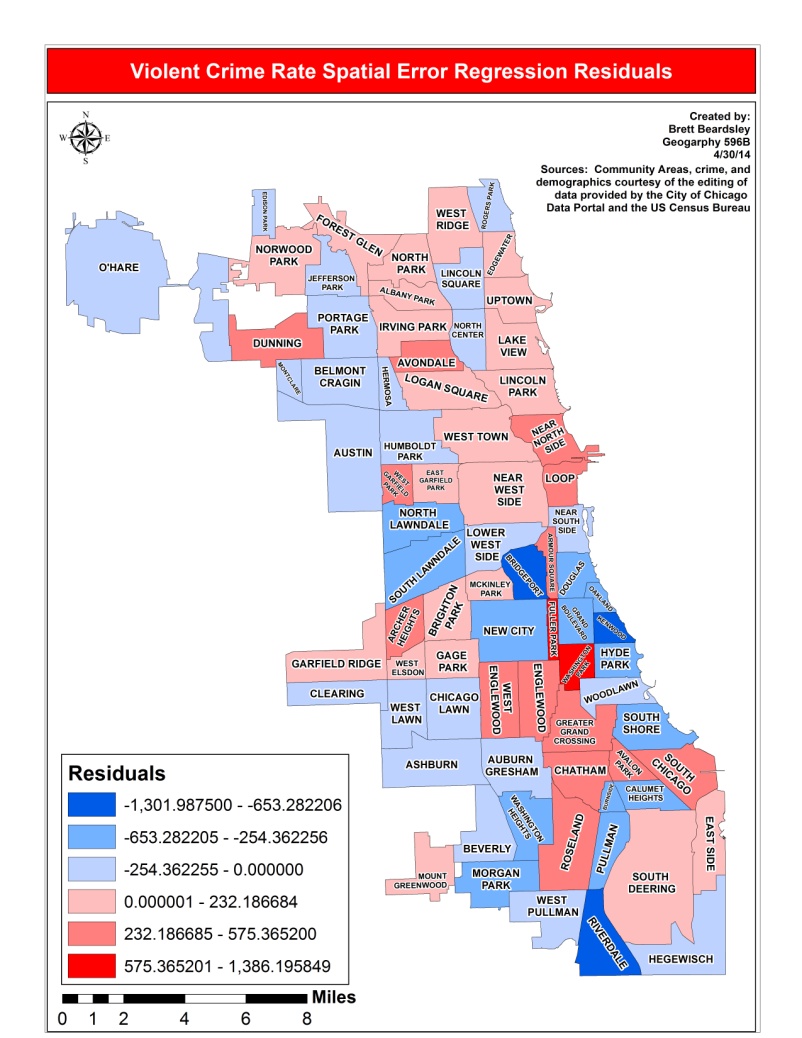
### Spatial Error Model-Violent Crime Rate

A spatial error model assumes that the residuals in neighboring Community Areas are related and similar for an unknown reason. It is basically assuming that there is some factor not in the model that is similarly affecting all Community Areas. In order to account for this spatial dependence a spatial error model adds a spatial error coefficient to the model. You can see the spatial error coefficient of 0.69 in Table 12.

**Table 12: Violent Crime Spatial Error Regression Results**

|  |  |  |  |
| --- | --- | --- | --- |
| Diagnostic | Coefficient | Value/Z-test | Probability |
| Constant | 1,728.25 | 5.25 | 0.000 |
| Per Capita Income | -0.02 | -2.97 | 0.003 |
| Percent White | -12.59 | -4.50 | 0.000 |
| Percent 15 to 24 | 11.49 | 0.81 | 0.415 |
| Percent Vacant | 58.36 | 6.71 | 0.000 |
| Percent Foreign | -26.02 | -5.76 | 0.000 |
| Spatial error term | 0.69 | 7.41 | 0.000 |
| Log likelihood |  | -559.81 |  |
| Akaike info criterion |  | 1,131.62 |  |
| Schwarz criterion |  | 1,145.68 |  |

The coefficient of each predictor variable in the spatial error model went farther away from zero compared to the OLS regression with the exception of percent vacant. Percent vacant still had the largest affect on violent crime with a coefficient of 58.36. The direction and significance of each predictor variable was the same in the error model as it was in the OLS model. The log likelihood, Akaike info criterion, and Schwarz criterion all moved closer to zero than their values in the OLS regression model, which is indicates a better goodness of fit for the spatial error regression model. The residual maps seen in Figure 6 for the OLS regression and spatial error maps also show that the error model has a better fit. You can see that there are less extreme low and high residuals in the spatial error residual map. There are also less values in the second highest and lowest residual categories in the spatial error residual map compared to the OLS residual map. This means the spatial error model more accurately predicts the total crime rate in Chicago Community Areas than the OLS model. The next model that was run was the OLS model for property crime.

**Figure 6: OLS and Spatial Error residuals**

### 

### OLS Regression-Property Crime Rate

The first OLS regression model for property crime had all nine predictor variables that made it through the correlation matrix reductions. This model had too much multicollinearity at 30.11, so another one had to be run. Mean household value was removed because it was insignificant and had strong and significant correlations with some of the remaining predictor variables. The multicollinearity in the second model was at a good level at 27.99, but there were still some variables that were insignificant to the property crime rate. Twenty five and over without a high school diploma, unemployed, and renters were removed because of non-significance and percentage of foreign born was removed because it was insignificant and heteroscedastic. This left four variables for the third model: per capita income, percentage of the population that is white, percentage of the population aged 15 to 24, and percentage of housing units that are vacant. The multicollinearity condition number went down again to 17.17 and every predictor other than population aged 15 to 24 was significant. Percentage of the population aged 15 to 24 was kept because it was the only variable that did not have strong autocorrelation. It also is a good variable to keep because crime is often committed by people in that age group. A satisfactory regression model for property crime had been reached after iterating through three models. The final model for property crime rate had the following variables: per capita income, percentage of the population that is white, percentage of the population aged 15 to 24, and percentage of housing units that are vacant.

Now that there is a final model for property crime rate I will go into a more specific analysis on more of the diagnostics from the regression model. Table 11 will illustrate many of the numbers that are going to be discussed. The r-squared value of 0.63 indicates that 63% of the variation in the property crime rate in Chicago

**Table 11: Property Crime OLS Regression Results**

|  |  |  |  |
| --- | --- | --- | --- |
| Diagnostic | Coefficient | Value/T-statistic | Probability |
| Constant | 398.53 | 0.34 | 0.732 |
| Per Capita Income | 0.09 | 6.00 | 0.000 |
| Percent White | -42.09 | -5.03 | 0.000 |
| Percent 15 to 24 | 100.74 | 1.43 | 0.158 |
| Percent Vacant | 158.30 | 4.39 | 0.000 |
| Multicollinearity Condition Number |  | 17.17 |  |
| Log likelihood |  | -673.71 |  |
| Akaike info criterion |  | 1,357.41 |  |
| Schwarz criterion |  | 1,369.13 |  |
| R-squared |  | 0.63 |  |
| F-statistic |  | 30.90 | 5.63e-15 |
| Moran’s I |  | 1.17 | 0.242 |
| Lagrange Multiplier (lag) |  | 3.95 | 0.047 |
| Robust LM (lag) |  | 0.40 | 0.012 |
| Lagrange Multiplier (error) |  | 2.71 | 0.529 |
| Robust LM (error) |  | 6.66 | 0.099 |

Community Areas was predicted or explained by the predictor variables in the model. That is a high percentage and shows that my model is a good one for predicting violent crime in a given Chicago Community Area. The F-statistic of 30.90 with a probability of 5.63e-15 rejects the null hypothesis that all regression coefficients are jointly zero. Both of these diagnostics show a good, goodness of fit for my model.

Percent vacant housing units had the largest coefficient at 158.30, which means that for every one percent vacant housing increases the property crime rate will increase by 158.30 crimes per 100,000 people. This also means that it has a larger affect on property crime than percent white and percent 15 to 24. As you can see in table 11 as per capita income, the percentage of people aged 15 to 24, and the percentage of vacant housing units increase, so does property crime. Therefore these variables have a positive relationship with property crime. The negative coefficient of percentage white indicates that as that the percentage of white people in a Chicago Community Area increases violent crime decreases. Therefore percent white has a negative relationship with property crime. The probabilities of the T-statistics indicate that all predictors other than percent aged 15 to 24 are significant. As was previously mentioned the multicollinearity number of 17.17 means that multicollinearity is not a problem. The insignificant Moran's I value means that autocorrelation is a not problem. The Lagrange Multiplier value and Robust LM value for lag are both significant at 0.047 and 0.012, which means the model needs a spatial lag regression specification. The Lagrange Multiplier and Robust LM values for error are insignificant at of 0.529 and 0.099.

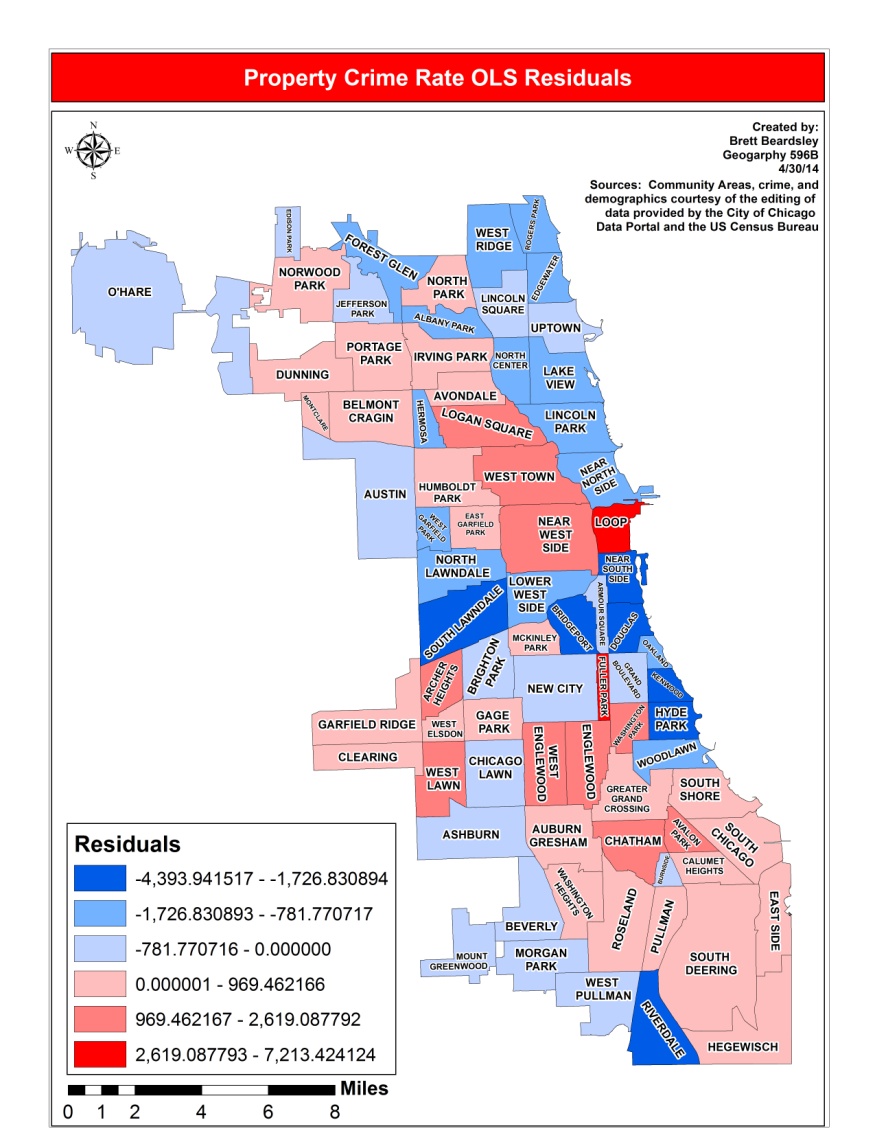
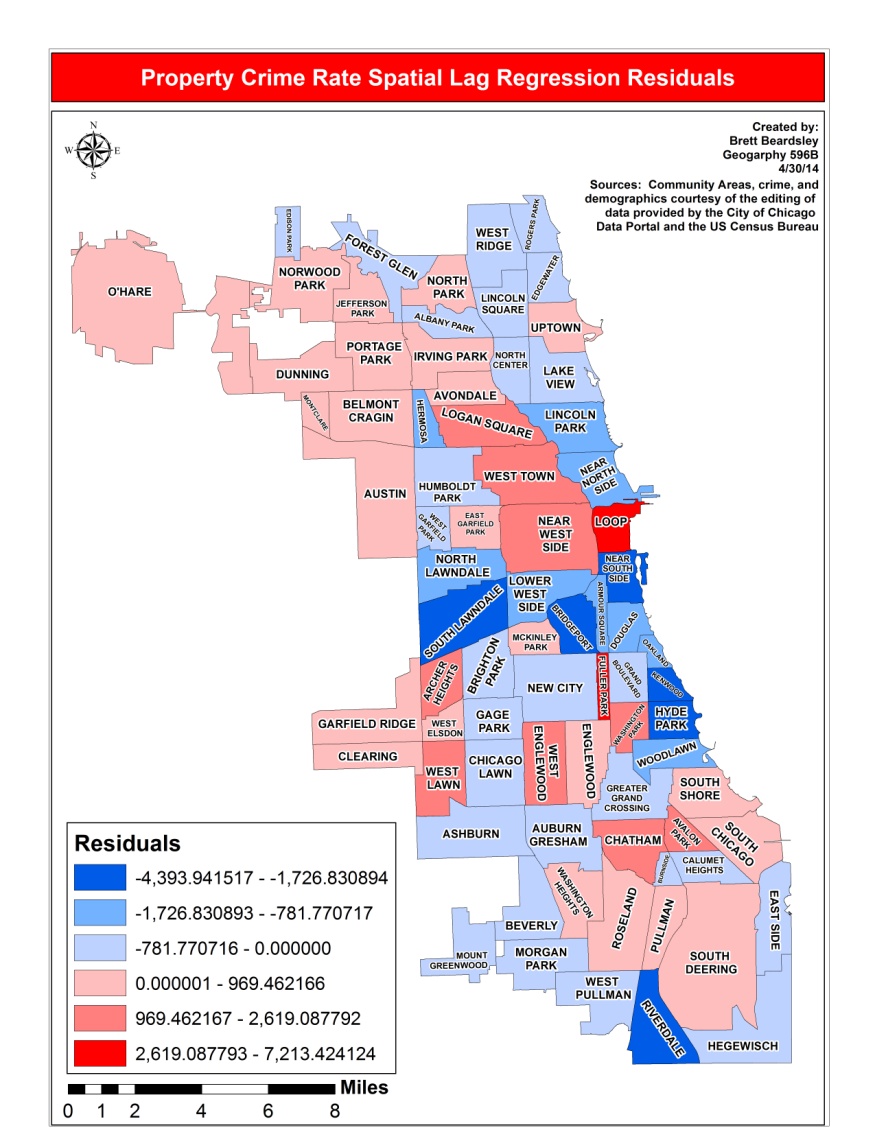
### Spatial Lag Model-Violent Crime Rate

Similar to the total crime rate OLS regression model, the property crime rate model called for a spatial lag model. The diagnostics can be seen in Table 12.

**Table 12: Property Crime Spatial Lag Regression Results**

|  |  |  |  |
| --- | --- | --- | --- |
| Diagnostic | Coefficient | Value/Z-test | Probability |
| Constant | -378.66 | -0.34 | 0.734 |
| Per Capita Income | 0.07 | 4.20 | 0.000 |
| Percent White | -29.25 | -3.24 | 0.001 |
| Percent 15 to 24 | 77.73 | 1.17 | 0.240 |
| Percent Vacant | 131.76 | 3.86 | 0.000 |
| Spatial lag term | 0.33 | 2.75 | 0.006 |
| Log likelihood |  | -671.23 |  |
| Akaike info criterion |  | 1,354.46 |  |
| Schwarz criterion |  | 1,368.52 |  |

The coefficient of each predictor variable in the spatial lag model decreased from the OLS regression model. You can see that vacancy had the largest affect on total crime rate with a coefficient of 131.76. The direction and significance of each is also the same as in the OLS regression. model. The values for the log likelihood, Akaike info criterion, and Schwarz criterion all became closer to zero from the OLS regression model, which indicates a better goodness of fit for the spatial lag regression model. Taking a look at the residual maps in Figure 7 between the two models also can help one visualize the better fit of the model. You can see that there are less extreme low and high residuals in the spatial lag residual map. There are also less values in the second highest and lowest residual categories in the spatial lag residual map compared to the OLS residual map. This means the spatial lag model more accurately predicts the property crime rate in Chicago Community Areas than the OLS model.

** **

**Figure 7: OLS and Spatial Lag residuals**

# Results and Conclusions

One of the research questions for this project was to see if there were any patterns and relationships among the statistically significant predictors across crime type. Creating the graph in Figure 8 that compares the predictor's coefficients across crime type is a good way to visualize the answer to that question. As you can see percent white and vacancy each had the same directional relationship across crime type. This graph also highlights how vacancy affects each crime type the most. Although you cannot see it because the values are so low, per capita income has a positive relationship with total and property crime and a negative relationship with violent crime. I believe that the reason for this is that property crime will usually only occur where there is something worth stealing. Therefore as per capita income increases there would theoretically be more and more items worth stealing. This is not the case with violent crime because violent crime does not have nearly as much to do with possessions. Overall these results give the idea that a great way to combat crime in a Chicago Community Area would be to eliminate as many vacant buildings as possible.

**Figure 8: Significant coefficients by crime type in OLS regression**

## Limitations

It is important to note the limitations of this study. One major limitation was the small number of observations of 77. This small number of observations eliminates the potential use of many types of analysis. Another limitation of this study is that the American Community Survey is only an estimate. It is not as complete as the 10 year censuses. This means that all of the predictors are not as accurately portrayed as they are in a census, which can skew the results. A third limitation on this study is that an index was not created for some of the similar predictors. This was done in some of the studies in the literature review and would have allowed for more predictors to have an influence on the models. For example instead of eliminating percent female headed households and percent households below poverty they could have been combined with other predictors to create an index value for all of them. This way the model would have been able to capture more predictors. This is similar to the violent and property crime indices for the outcome variables. As was previously mentioned the rape statistics were skewed somewhat when comparing Chicago to the rest of the country because Chicago's definition of rape allows for more rapes to be counted. In future studies this would not be a problem because in 2012 the federal government changed their rape definitions, which would make them similar to Chicago's definition (Savage, 2012). However since all rapes in Chicago were classified the same this did not have an effect on the study. It is just something to consider if someone were to compare this study to a similar one in a different city.

## Further Studies

If someone were to do this study again there are numerous ideas to consider. One part that should be done differently is to use more up to date 5-year ACS data. When this study was conducted the most up to data 5-year ACS was the 2007 to 2011 version. A new one comes out every year and 2008 to 2012 is already out, so whenever someone else would want to conduct a similar study they should use the most up to date ACS. As was mentioned in the Limitations section, the predictors would ideally be indexed in order to get the most out of as many predictors as advisable. Doing a similar study with a different unit of analysis would also shed some new light on some areas. There are 805 census tracts in Chicago and more census blocks than that. Using tracts or blocks would present challenges as well. There would be more variability across units of analysis.  When using tracts or blocks you would be using such a small unit of analysis that you could question the sample size of the tract or block.  There would also be many more outliers.  You could have one block that is just absolutely horrible and a block next to it that still is not a safe place, but for some reason not a lot of crime happens there.  This would be misleading and would skew the overall results.

Another great idea would be to take a deeper look at the residuals of the Community Areas and perform a more in depth analysis of why the models' predictions are far off. One could potentially concentrate on one Community Area and write a complete analysis on just that area. The amount of possible further study when looking at residuals is bountiful to say the least. Being a Chicago native it would also be interesting to examine why Fuller Park was at the top of so many crime categories. It is not a notoriously dangerous neighborhood and to see it at the top of multiple crime types was very surprising. There has to be a reason for its infamous place in this study. Lastly, the crime data came in point form, so a point pattern analysis of these crimes would be very doable and useful.

# 

# Works Cited

Adams, D.R. (2007). *Introduction to Regression.* Retrieved 04 20, 2014 from Princeton.edu: http://dss.princeton.edu/online\_help/analysis/regression\_intro.htm

Anselin, L. (2005). *GeoDa Workbook.* Retrieved 04 29, 2014 from GeoDa: https://geodacenter.asu.edu/system/files/geodaworkbook.pdf

Arnio, A. N. & Baumer, E. P. (2012). Demography, foreclosure, and crime: Assessing spatial heterogeneity in contemporary models of neighborhood crime rates. *Demographic Research 26:18,* 449-488.

Berg, M.T., Brunson, R.K., Stewart, E.A., & Simons, R.L. (2011). Neighborhood Cultural Heterogeneity and Adolescent Violence. *Journal of Quantitative Criminology 28,* 411- 435.

Boggs, S. (1965). Urban Crime Patterns. *American Sociological Review 30:6.* 899-908.

Bowers, K. & Hirschfield, A. (1999). Exploring links between crime and disadvantage in north-west England: an analysis using geographical information systems. *International Journal of Geographical Information Science 13:2.* 159-184.

Ceccato, V. (2005). Homicide in Sao Paulo, Brazil: Assessing spatial-temporal and weather variations. *Journal of Environmental Psychology 25:3,* 307-321.

Census. (2014). *State & County QuickFacts.* Retrieved 03 26, 2014, from Census.gov: http://quickfacts.census.gov/qfd/states/17/1714000.html

Census. (2014). *What is the American Community Survey.* Retrieved 04 03, 2014 from Census.gov: https://www.census.gov/acs/www/

City of Chicago. (2014). *Data Portal*. Retrieved 10 31, 2013, from data.cityofchicago.org: https://data.cityofchicago.org/Public-Safety/Crimes-2001-to-present/ijzp-q8t2

Earls, F., Morenoff, J.D, & Sampson, R.J. (1999). Beyond Social Capital: Spatial Dynamics of Collective Efficacy for Children. *American Sociological Review 64:5,* 633-660.

ESRI. (n.d). ArcGIS 10 Help. Redlands, CA.

Federal Bureau of Investigation. (2011). *Uniform Crime Reports*. Retrieved 03 24, 2014, from Fbi.gov: [http://www.fbi.gov/about-us/cjis/ucr/crime-in-the-u.s/2011/crime-in-the-u.s.- 2011/tables/table\_12\_crime\_trends\_by\_population\_group\_2010-2011.xls](http://www.census.gov/acs/www/)

Goode, E.. (2011, 09 28). *Rape Definition Too Narrow in Federal Statistics, Critics Say.* Retrieved 04 08, 2014, from Nytimes.com: http://www.nytimes.com/2011/09/29/us/federal-rules-on-rape-statistics-criticized.html?pagewanted=all&\_r=0

Huffington Post. (2013, 09 12). 'Chi-Raq' Documentary: *Will Robson-Scott Highlights Shooters, Survivors or City Violence.* Retrieved 03 25, 2014, from Huffingtonpost.com: http://www.huffingtonpost.com/2013/09/12/chi-raq-documentary-\_n\_3916444.htm

Graif, C. & Sampson, R. J. (2009). Spatial Heterogeneity in the Effects of Immigration and Diversity on Neighborhood Homicide Rates. *Homicide Studies* *13:3,* 242-260.

Matthews, S.A., Yang T-C., Hayslett, K.L., & Ruback, R.B. (2010). Built environment and property crime in Seattle, 1998-2000: a Bayesian analysis.  *Environment and Planning 42:6,* 1403-1420.

Morenoff, J.D. (2003). Neighborhood Mechanisms and the Spatial Dynamics of Birth Weight. *American Journal of Sociology 108:5,* 976-1017.

O'Sullivan, D. and Unwin D. (2010). *Geographic Information Analysis.* Hoboken: John Wiley & Sons, Inc.

Peterson, N. (2012, 05 23). *Chicago's murder rate mirrors warzone, federal data shows.* Retrieved 03 26, 2014, from Austinweeklynews.com: http://www.austinweeklynews.com/News/Articles/5-23-2012/Chicago's-murder-rate- mirrors-war-zone,-federal-data-shows/

Raudenbush, S.W., Sampson, R.J., & Sharkey, P. (2008). Durable effects of concentrated disadvantage on verbal ability among African-American children. *Proceedings of the National Academy of Sciences* *105:3,* 845-852.

Savage, C. (2012, 01 06). *U.S. to Expand Its Definition of Rape in Statistics.* Retrieved 05 11, 2012, from Nytimes.com: http://www.nytimes.com/2012/01/07/us/politics/federal-crime- statistics-to-expand-rape-definition.html?\_r=0

Shaw, C.R. (1929). *Delinquency Areas.* Chicago: University of Chicago Press.

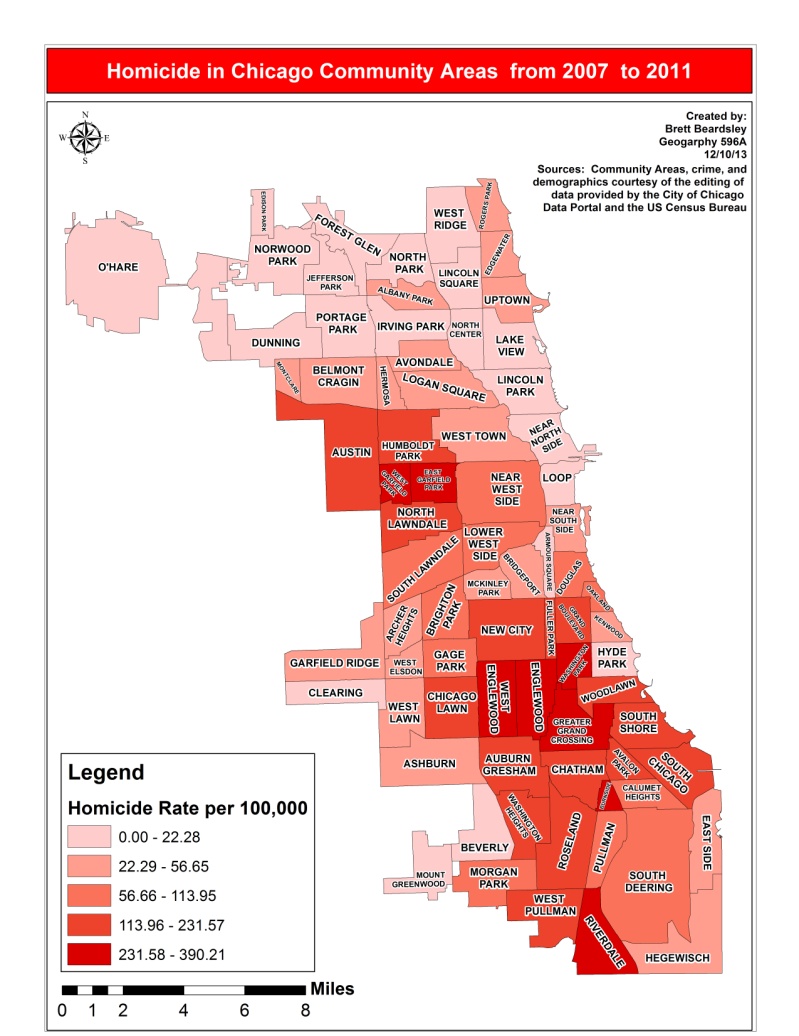
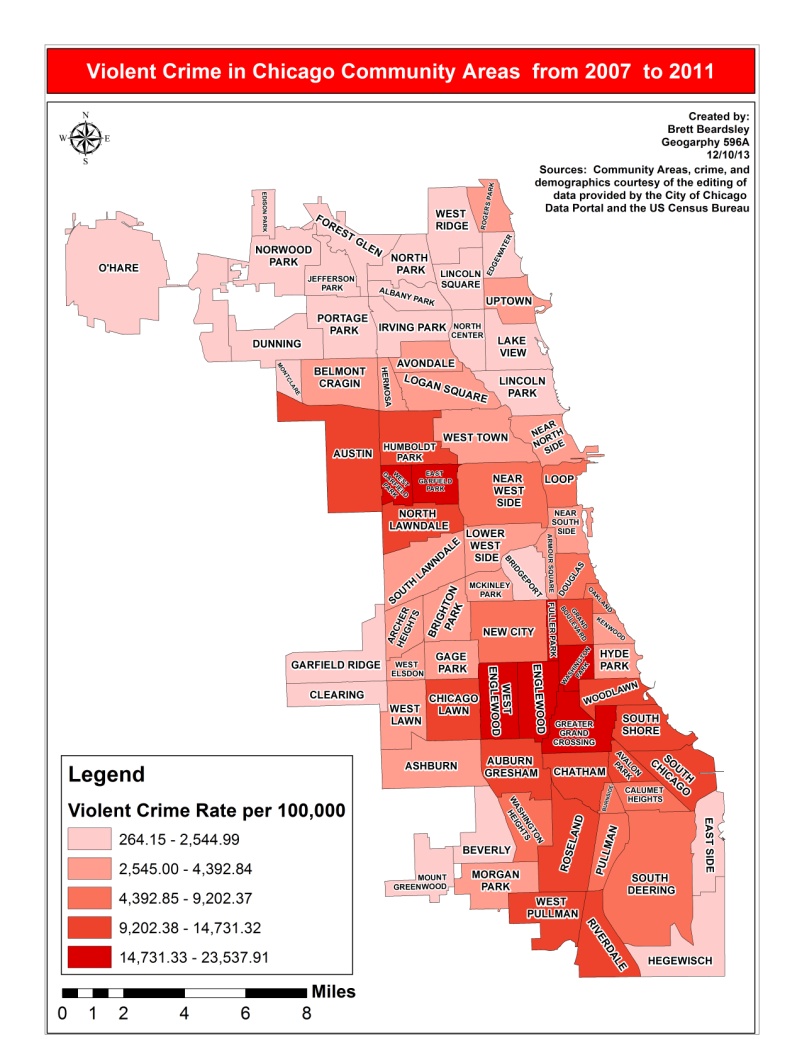
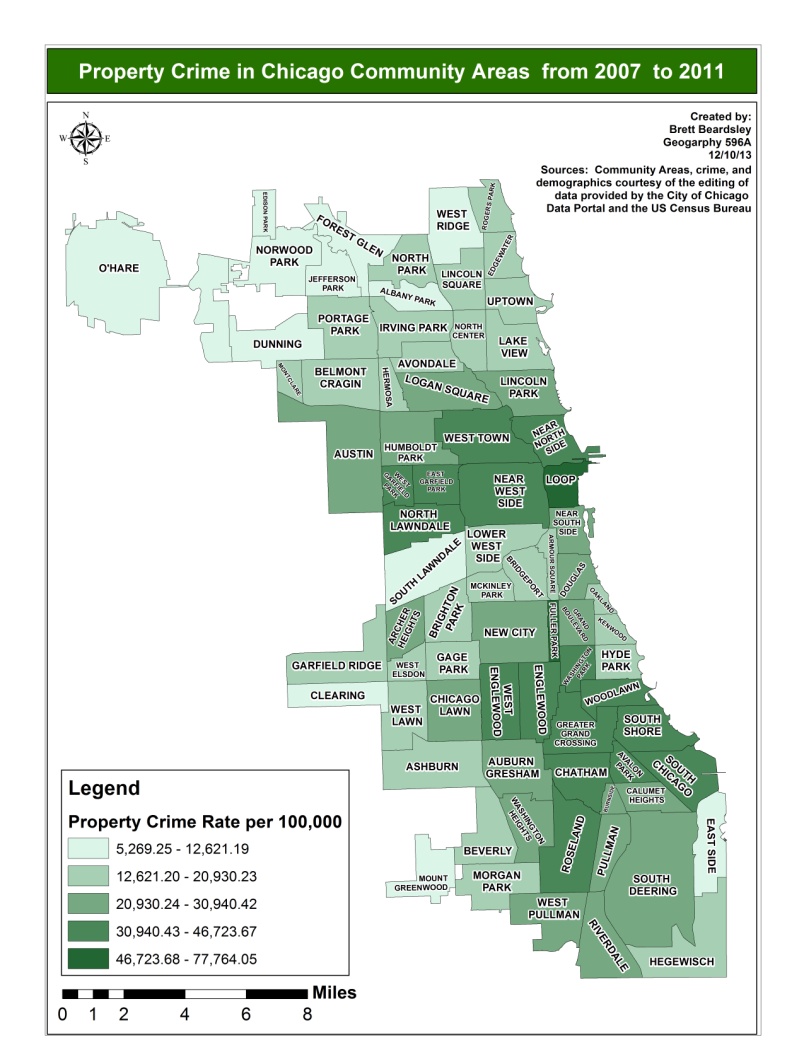
Stockburger, D.W. (1998). *Introductory Statistics: Concepts, Models, and Applications.* Retrieved 04 15, 2014 from Missouristate.edu: http://www.psychstat.missouristate.edu/introbook/sbk17m.htm

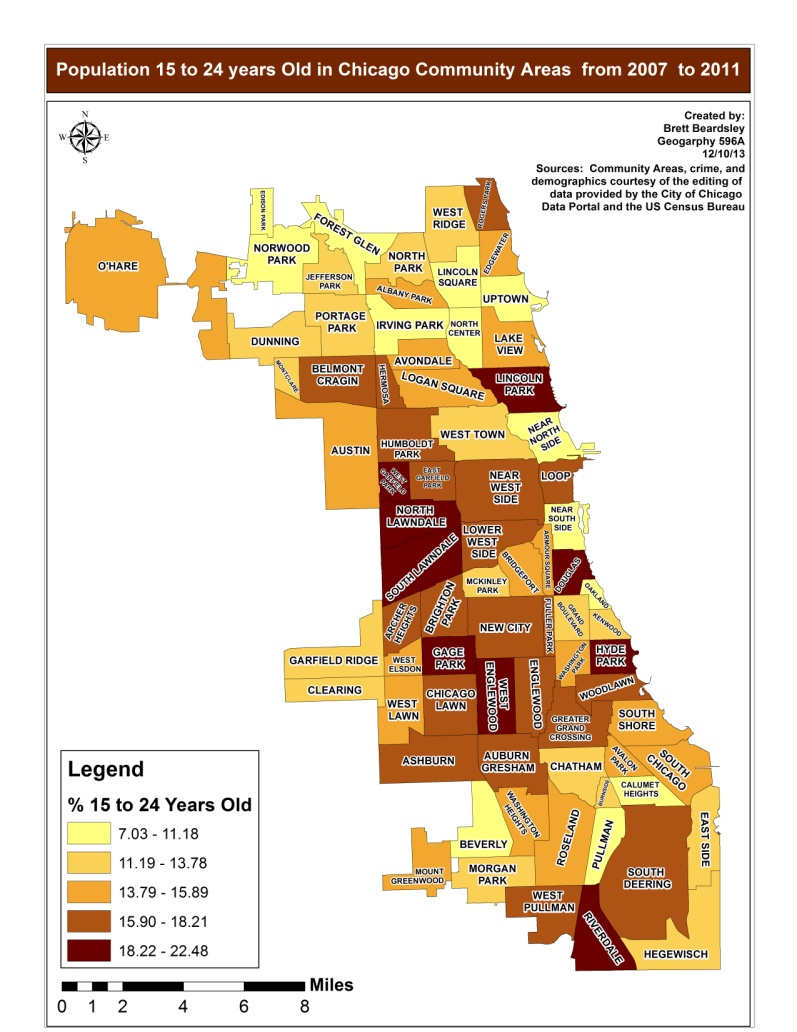
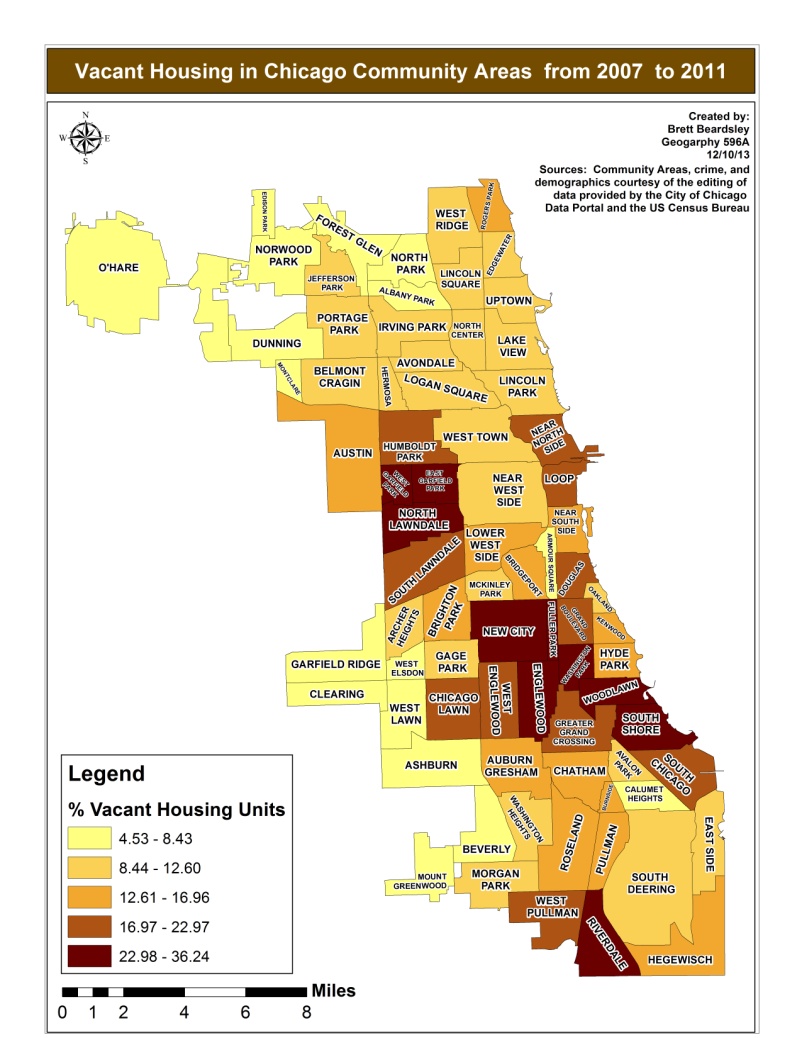
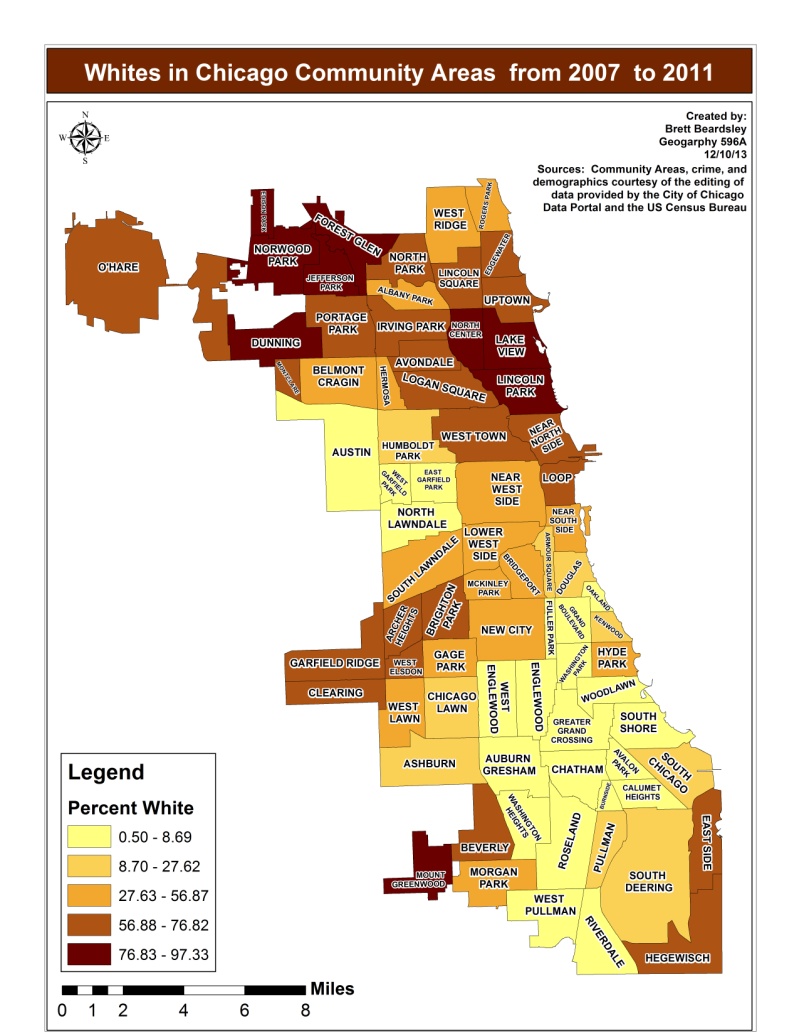
Sykes, A.O. (1992). *An Introduction to Regression Analysis.* Retrieved 04 19, 2014 from UChicago.edu: http://www.law.uchicago.edu/files/files/20.Sykes\_.Regression.pdf

White, R.C. (1932). The Relation to Felonies to Environmental Factors in Indianapolis. Social Forces 10:4, 498-509.

This document is published in fulfillment of an assignment by a student enrolled in an educational offering of The Pennsylvania State University. The student, named above, retains all rights to the document and responsibility for its accuracy and originality.

# Appendix A: Selected Choropleth Maps of Variables and Total Crime Rate Chart

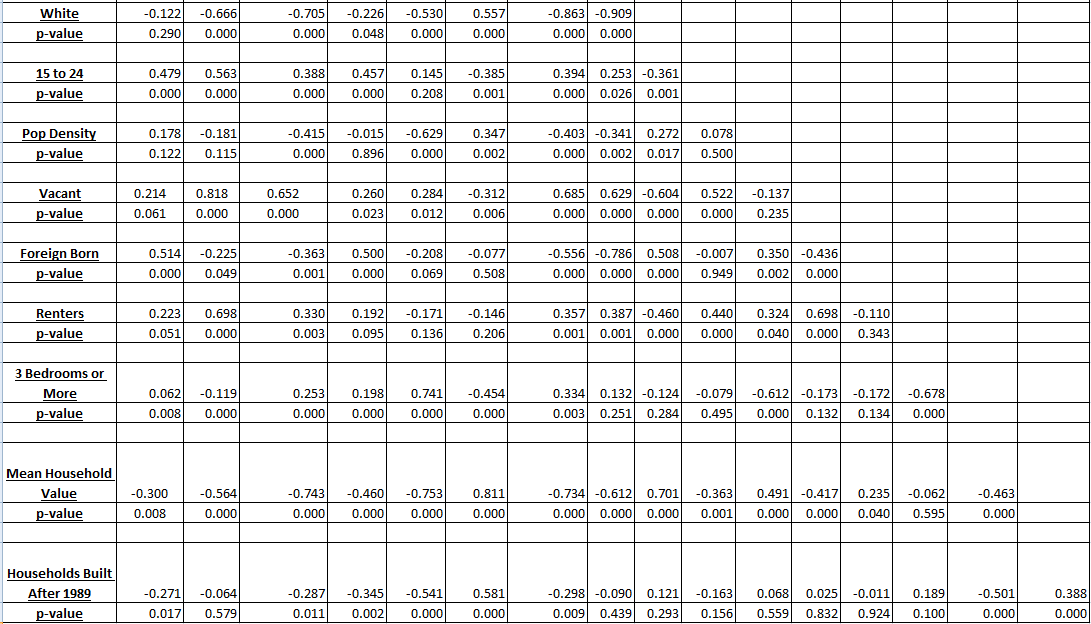
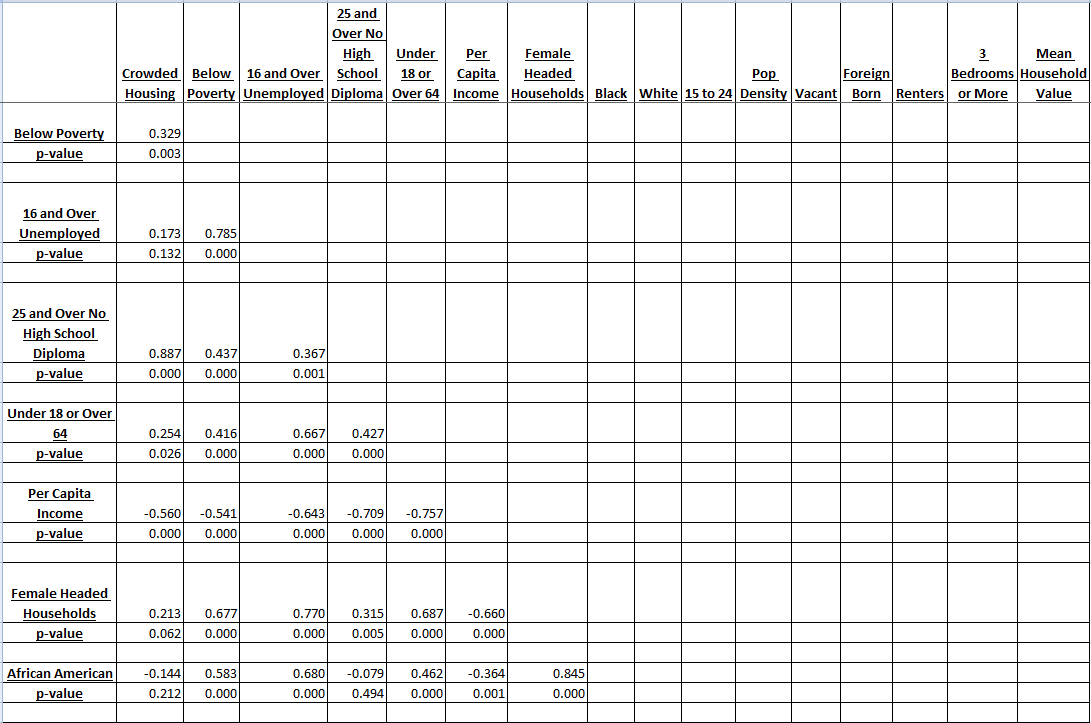




|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Community Area | Total Crime Rate | Rank | Community Area | Total Crime Rate | Rank |
| Fuller Park | 16,815.88 | 1 | Lake View | 4,373.55 | 43 |
| Loop | 16,764.65 | 2 | Avondale | 4,176.72 | 44 |
| Washington Park | 13,584.91 | 3 | Armour Square | 4,152.93 | 45 |
| Englewood | 13,436.17 | 4 | Belmont Cragin | 4,081.77 | 46 |
| West Englewood | 11,643.74 | 5 | Gage Park | 3,889.22 | 47 |
| East Garfield Park | 10,890.22 | 6 | Uptown | 3,818.53 | 48 |
| Greater Grand Crossing | 10,810.95 | 7 | Rogers Park | 3,814.52 | 49 |
| West Garfield Park | 10,717.15 | 8 | Lower West Side | 3,785.27 | 50 |
| Chatham | 10,583.81 | 9 | Irving Park | 3,726.40 | 51 |
| South Shore | 10,499.63 | 10 | Ashburn | 3,714.95 | 52 |
| Near West Side | 9,960.30 | 11 | North Center | 3,582.61 | 53 |
| Woodlawn | 9,502.20 | 12 | Hermosa | 3,509.37 | 54 |
| Avalon Park | 9,497.61 | 13 | North Park | 3,454.38 | 55 |
| North Lawndale | 9,434.74 | 14 | Brighton Park | 3,445.91 | 56 |
| South Chicago | 9,011.43 | 15 | West Elsdon | 3,406.08 | 57 |
| Roseland | 8,898.36 | 16 | South Lawndale | 3,353.11 | 58 |
| Near North Shore | 8,791.24 | 17 | Garfield Ridge | 3,322.23 | 59 |
| Grand Boulevard | 8,422.47 | 18 | Montclare | 3,295.45 | 60 |
| Auburn Gresham | 8,307.35 | 19 | Hegewisch | 3,259.50 | 61 |
| Austin | 7,818.88 | 20 | Beverly | 3,157.55 | 62 |
| West Pullman | 7,709.72 | 21 | Portage Park | 3,145.91 | 63 |
| Chicago Lawn | 7,503.24 | 22 | Bridgeport | 3,088.01 | 64 |
| West Town | 7,357.34 | 23 | Edgewater | 3,034.17 | 65 |
| Humboldt Park | 7,252.67 | 24 | Lincoln Square | 2,998.86 | 66 |
| Riverdale | 7,065.78 | 25 | West Ridge | 2,849.52 | 67 |
| Washington Heights | 7,064.75 | 26 | East Side | 2,687.21 | 68 |
| New City | 6,988.88 | 27 | Albany Park | 2,583.86 | 69 |
| Calumet Heights | 6,748.32 | 28 | Clearing | 2,561.63 | 70 |
| Burnside | 6,644.93 | 29 | Dunning | 2,499.64 | 71 |
| Pullman | 6,558.88 | 30 | Jefferson Park | 2,091.27 | 72 |
| Douglas | 6,469.62 | 31 | O'Hare | 1,779.45 | 73 |
| South Deering | 6,186.88 | 32 | Forest Glen | 1,660.65 | 74 |
| Lincoln Park | 5,846.15 | 33 | Mount Greenwood | 1,658.67 | 75 |
| Logan Square | 5,693.58 | 34 | Norwood Park | 1,515.59 | 76 |
| Near South Side | 5,622.03 | 35 | Edison Park | 1,128.62 | 77 |
| Archer Heights | 5,503.53 | 36 |
| Oakland | 5,276.39 | 37 |
| Hyde Park | 4,893.55 | 38 |
| Morgan Park | 4,740.95 | 39 |
| Kenwood | 4,686.18 | 40 |
| West Lawn | 4,553.16 | 41 |
| McKinley Park | 4,483.43 | 42 |

# Appendix B: Total Crime Rate Chart and Correlation Matrices

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Crime | Murder and Non Negligent Manslaughter | Aggravated Assault | Criminal Sexual Assault | Robbery | Burglary | Motor Vehicle Theft | Arson |
| Aggravated Assault | 0.943 |  |  |  |  |  |  |
| p-value | 0.000 |  |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| Criminal Sexual Assault | 0.934 | 0.967 |  |  |  |  |  |
| p-value | 0.000 | 0.000 |  |  |  |  |  |
|  |  |  |  |  |  |  |  |
| Robbery | 0.863 | 0.936 | 0.930 |  |  |  |  |
| p-value | 0.000 | 0.000 | 0.000 |  |  |  |  |
|  |  |  |  |  |  |  |  |
| Burglary | 0.839 | 0.857 | 0.899 | 0.819 |  |  |  |
| p-value | 0.000 | 0.000 | 0.000 | 0.000 |  |  |  |
|  |  |  |  |  |  |  |  |
| Motor Vehicle Theft | 0.849 | 0.902 | 0.882 | 0.922 | 0.823 |  |  |
| p-value | 0.000 | 0.000 | 0.000 | 0.000 |  |  |  |
|  |  |  |  |  |  |  |  |
| Arson | 0.829 | 0.795 | 0.757 | 0.676 | 0.753 | 0.757 |  |
| p-value | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 | 0.000 |  |
|  |  |  |  |  |  |  |  |
| Larceny/Theft | 0.297 | 0.401 | 0.434 | 0.567 | 0.283 | 0.445 | 0.153 |
| p-value | 0.009 | 0.000 | 0.000 | 0.000 | 0..013 | 0.000 | 0.184 |



|  |  |  |  |
| --- | --- | --- | --- |
| Variables | Total Crime | Violent Crime | Property Crime |
| Crowded Housing | -0.069 | 0.090 | -0.137 |
| p-value | 0.549 | 0.435 | 0.236 |
|  |  |  |  |
| 16 and Over Unemployed | 0.544 | 0.756 | 0.404 |
| p-value | 0.000 | 0.000 | 0.000 |
|  |  |  |  |
| 25 and Over No High School Diploma | -0.005 | 0.186 | -0.091 |
| p-value | 0.967 | 0.106 | 0.431 |
|  |  |  |  |
| Under 18 or Over 64 | 0.118 | 0.446 | -0.041 |
| p-value | 0.308 | 0.000 | 0.722 |
|  |  |  |  |
| Per Capita Income | -0.097 | -0.437 | 0.066 |
| p-value | 0.400 | 0.000 | 0.571 |
|  |  |  |  |
| White | -0.654 | -0.816 | -0.527 |
| p-value | 0.000 | 0.000 | 0.000 |
|  |  |  |  |
| 15 to 24 | 0.388 | 0.440 | 0.332 |
| p-value | 0.000 | 0.000 | 0.003 |
|  |  |  |  |
| Population Density | -0.159 | -0.254 | -0.103 |
| p-value | 0.168 | 0.026 | 0.375 |
|  |  |  |  |
| Vacant | 0.747 | 0.836 | 0.647 |
| p-value | 0.000 | 0.000 | 0.000 |
|  |  |  |  |
| Foreign Born | -0.589 | -0.622 | -0.526 |
| p-value | 0.000 | 0.000 | 0.000 |
|  |  |  |  |
| Renters | 0.502 | 0.533 | 0.447 |
| p-value | 0.000 | 0.000 | 0.000 |
|  |  |  |  |
| 3 Bedrooms or More | -0.167 | -0.049 | -0.252 |
| p-value | 0.147 | 0.671 | 0.027 |
|  |  |  |  |
| Mean Household Value | -0.344 | -0.579 | -0.210 |
| p-value | 0.002 | 0.000 | 0.067 |
|  |  |  |  |
| Households Built after 1989 | 0.217 | -0.098 | 0.343 |
| p-value | 0.058 | 0.398 | 0.002 |

This document is published in fulfillment of an assignment by a student enrolled in an educational offering of The Pennsylvania State University. The student, named above, retains all rights to the document and responsibility for its accuracy and originality.