VISUALIZATION AND MODELING FOR CRIME DATA INDEXED BY ROAD SEGMENTS

by

Krista Heim

A Dissertation
Submitted to the
Graduate Faculty
of
George Mason University
in Partial Fulfillment of
The Requirements for the Degree
of
Doctor of Philosophy
Statistical Science

Committee:

_____	Dr. Daniel Carr, Dissertation Director
_____	Dr. Cynthia Lum, Committee Member
_____	Dr. Liansheng Tang, Committee Member
_____	Dr. Clifton Sutton, Committee Member
_____	Dr. William Rosenberger, Department Chair
_____	Dr. Kenneth S. Ball, Dean, Volgenau School of Engineering
Date: _____	Summer Semester 2014 George Mason University Fairfax, VA

Visualization and Modeling for Crime Data Indexed by Road Segments

A Dissertation submitted in partial fulfillment of the requirements for the degree of Doctor of Philosophy at George Mason University

By

Krista Heim
Master of Science
George Mason University, 2011
Bachelor of Science
Arcadia University, 2009

Director: Daniel Carr, Professor
Department of Statistical Science

Summer Semester 2014
George Mason University
Fairfax, VA

DEDICATION

This dissertation is dedicated to my parents, friends, and my academic advisor, Dr. Daniel Carr, for all of their support and encouragement.

ACKNOWLEDGEMENTS

I would like to thank my parents for giving me the confidence to believe I can do anything I set my mind to, and for their support throughout the journey to getting my PhD. I would especially like to thank my advisor, Dr. Carr, for his guidance throughout my research. He has been very encouraging throughout my studies and helped me understand the importance of statistical visualization with his extensive background in the field.

I want to thank George Mason's Center for Evidence-Based Crime Policy that invited me to a very informative conference that provided much insight into criminology. Especially I want to thank Dr. Cynthia Lum, who guided me to the literature on the criminology of place and to the Alexandria Police Department, who provided data and related explanations. I would also like to thank my other committee members, Dr. Clifton Sutton and Dr. Liansheng Tang for their support.

Thanks to the Alexandria Police Department, specifically to Captain Chris Wemple and Matt Smith, for providing me with crime data and police calls for service data. Their help in understanding the data has been invaluable. Also thanks Dr. Edward Zolnik and Jeanette Chapman of the GMU Center for Regional Analysis for providing me with some property value data.

I'm also thankful to Dr. Chris Saunders, who provided some of my research assistant funding under an NIJ forensics grant and the GMU Volgenau School of Engineering that provided a semester of funding toward finishing my dissertation.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF EQUATIONS

ABSTRACT

VISUALIZATION AND MODELING FOR CRIME DATA INDEXED BY ROAD SEGMENTS

Krista Heim, Ph.D.

George Mason University, 2014

Dissertation Director: Dr. Daniel Carr

This research develops crime hotspot analysis and visualization methodology that use street segments as the basic study unit. This incorporates the distance between points along a polyline rather than the standard Euclidean distance and has some distinct advantages over past methods. For each crime, this method creates a weight according to its distance from each road segment of its surrounding block. To create the hotspot visualization map, crime counts are smoothed over road segments based on the distance to nearest segments and the angle at which nearest roads meet at intersections. Crime data from the City of Alexandria, VA Police Department and San Francisco, CA (available at data.sfgov.org) are considered here using a combination of conventional ArcGIS and R graphics.

I assume that demographic variables related to crime in large areas are still relevant to crime rates at the local level and seek to make use of the most spatially detailed data accessible. Decennial demographic variables at the block level for 2010 from the U.S. Census are associated with road segments by assigning the available values to the surrounding segments of each block. These variables include age, gender, population, and housing for both locations. Variables also considered are police calls for service, housing prices, elevation and speed limits.

I discuss/compare area crime counts with polyline crime counts using (zero-inflated) Poisson and Negative Binomial regression with crime-related covariates, as well as MCMC Poisson-Gamma Conditional Autoregressive (CAR) model in CrimeStat IV and a localized CAR model in R using distances between segments as weights. Conditional variable importance is measured using conditional random forest modeling to see which of the covariates are the most important predictors of crime and to decide which variables are the most appropriate to consider for visualization. Principal components are also used to create independent linear combinations of predictor variables. While most visualization approaches for street segments have emphasized one variable at a time, this research uses a 3 x 3 grid of maps using DPnet to highlight each grouping of road segments associated with classes based on two covariates. This multivariate visualization will allow us to explore multiple variables at a time and their patterns along a road network.

CHAPTER 1. INTRODUCTION

1.1 Problem Statement

This dissertation develops new techniques of visualization, smoothing, and modeling over road networks, with emphasis on crime data. While recent research in the "criminology of place" has led to deeper understanding of relating local patterns to streets, there remain analytic and visualization challenges to address. The basic notion here is that distance along street segments is sometimes more relevant to the understanding of crime concentration than great arc distance over the entire space. Analyses that use larger spatial units, such as census tracts, as opposed to street segments may hide the variability of crime rates among the streets within those areal units. The importance of streets makes sense in terms of human activity and thinking. Examples include police patrolling, emergency response, logistics in moving from place to place, and territorial boundaries as defined by city zoning or gangs. Recent advances have called attention to the crime patterns that appear on street segments, suggesting that streets can provide a useful geospatial foundation for analysis of selected types of crime data.

The modeling and visualization of geospatial data associated with lines provides perspectives that are particularly relevant to phenomena involving pipelines, streams, and streets. Both the phenomena and the audience that seeks to understand and interact with

them influence the perspectives that are useful. The criminology of place research (Weisburd et al, 2012) calls attention to crime pattern on street segments. Street segments are represented on a map as either a single line or polyline between two intersections. A polyline is a connected series of line segments; for convenience, I use the term polyline to refer to the geometric representation of the street segment, even when there is only one line. Many kinds of crime occur in close proximity to street segments. Streets are important for vehicle and foot transportation. In terms of city design and regulation, streets bound industrial use and housing zones. Locations along streets are important for many businesses. Streets can bound areas associated with different kinds of social activity and demographics compositions. Street-based perspectives are not only useful for their increasingly understood connection to crime, but also because of their multifaceted roles in urban dwellers thought processes.

However, the modeling and visualization of data associated with lines and polylines that represent street segments is less common than the modeling and visualization of data associated with points and areas. Correspondingly, the methodology is less mature. For lines and polylines there are fewer published examples and gaps to address for the combination of geospatial structure, phenomena, and human cognition, motivating adapting or creating more suitable methods. I provide one approach for developing and illustrating methodology that converts point and area statistics into polyline statistics (including Census block data), providing a unified framework for spatially detailed regression modeling.

The problems I address include:

- Gathering crime relevant data,
- Converting potentially crime-relevant statistics indexed by areas and points to statistics indexed by street segments (lines or polylines) for use in regression models and graphics,
- Developing smoothing methods for street segment statistics based on alternative notions of street segment proximity,
- Building regression models to identify important variables and make predictions,
- Creating graphics and visualization methodology to show observed data and estimated values to support communication, data exploration and model criticism.

I associate statistics (such as crime counts) with roads based on an inverse distance weighting function involving distance between each crime point and its surrounding segment midpoints. It is intuitive that each surrounding segment should be given some value of the crime because that crime is contained within those segments. The motivation for the weighting is in terms of accessibility to the crime location. A road segment is considered more relevant to the crime if it is the closer and thus receives more value.

For visualization, my unique smoothing algorithm uses a weighting function based on distance between nearest street segments at two levels and the angle at which these streets meet at intersections. The basic idea of smoothing is that it uses an

averaging of statistics with neighboring statistics (in this case crime counts on neighboring segments) to cancel out some of the noise and help reveal more of the underlying structure. Behaviors on streets are not completely independent of each other, but are also influenced by streets in surrounding areas. Smoothing reflects this phenomena and helps to pick out clusters of high crime streets more clearly. Police would rather focus on a series of connected roads to patrol rather than just one individual block at a time. Smoothing will help distinguish the groupings of roads to focus attention.

This algorithm is a new contribution to the literature on smoothing. Segments that are a far distance away will get less weight because they will be less relevant to the original segment in terms of the criminology of place. Segments within direct view of the original segments will get higher weights, as being able to look down that street to the next street makes the two more relatable and relevant to each other. Segments with sharp intersections that make small angles will also have high weight. There is easier access between the two segments, with ability to move through yards and alleyways. Segments at right angles then have the least value because there is no easy visibility or access between the two streets as seen previously. Also, studies in Seattle showed that in some areas, streets in one direction fell in different clusters than streets crossing them orthogonally (Weisburd et. al 2012). While such smoothing has appeal due to focus on crime, it does not directly adjust for additional variables related to crime.

Using data that wasn't originally assigned to streets segments will help see patterns with crime and related variables that may not have been possible to before on roads. Previous methods did not incorporate block data, but block data can still have pertinent information that describes the demographics of the area, that could relate to why the crime is high or low in those areas. Blocks are bordered by road segments, so assigning segments these block values seems intuitive.

In terms of modeling, my hypothesis is that street segment-based analysis of crime is better than area-based analysis, better fitting the overall crime data and predicting future crime. I compare modeling of crime counts with related covariates over polylines (street segments) to typical modeling of crime counts over area. In this research I borrow from advances made in the modeling of point and area data. I explore several different modeling methods that represent different facets of the data. Past parametric modeling based on discrete event data, such as cancer deaths, were sometimes founded on the use of the Poisson distribution. However, the distribution has only one parameter that needs to account for both the mean and variance in the data. The presence of over-dispersion (excess variance) in the data motivates the use of more appropriate models. One such discrete event model is the negative binomial model, which has two parameters that determine mean and variance. The negative binomial model of road segment crime counts fits better than the Poisson model. Counts of zeros also often appear in area-based count data, and many Alexandria, VA road segments have zero crimes. Researchers have developed zero-inflated models to deal with this issue, and using this type of model yields still better fit values.

Geospatial models often address spatial correlation; the last parametric model elaboration is the hierarchical CAR Bayes Model, which addresses the spatial dependence of the data. I take a novel approach to the CAR model by measuring proximity based on neighboring road segments rather than neighboring polygons. I use both the area-based and road-based models to also predict known crime values. Specifically, I model 2006-2008 data and compare fitted values with the observed 2009 data. I find that for the example of Alexandria, VA, I could better predict crime over road segments using my model than when they were aggregated over area units. The small size of streets reduces spatial heterogeneity, leading to smaller errors of prediction. This could be helpful in the field of predictive policing.

Multicollinearity becomes an issue in the models discussed above, with many predictors highly related to one another. This makes the individual interpretations of the significance of each variable unreliable. Conditional random forests and principal components analysis are both explored to select which covariates in the model are the strongest predictor variables. Using selected variables (or linear combinations of them), I create a multivariate visualization using DPnet that includes the crime counts with two crime-related variables at a time. Multivariate visualization of important variables with crime enables us to see in what way patterns of crime in different areas are influenced by multiple variables at once.

1.2 Organization

Chapter 2 provides background on related research in the field of criminology. Fields in social sciences have often developed bodies of knowledge by defining concepts related to the phenomena of interest and by thinking about relationships among the concepts. The presence of quantitative variables associated with the concepts opens the door to building empirical models. Scholarly research in criminology is often guided by this body of knowledge and discussion of empirical models makes connections to this body of knowledge. This chapter also includes background on statistical methods related to this research, addressing hotspot detection and smoothing methodology.

Chapter 3 will present the crime data from Alexandria, VA and San Francisco, CA and give some exploratory data analysis. For Alexandria, VA I will focus on both the full crime data set and a subset of assault crimes. This chapter will also explore the crime-related variables obtained to model with the crime data, along with some criminological motivation for using them.

Chapter 4 will go into detail on the smoothing algorithm that I created. It will discuss both the assignment point and area data to street segments and smoothing crime counts over street segments. Note that I will use the terms “roads” and “streets” interchangeably. Smoothing visualization results are given for Alexandria and San Francisco, CA.

Chapter 5 discusses regression modeling. I discuss modeling crime counts with relevant covariates using Poisson and Negative Binomial regression, along with the zero-inflated versions of these models. Following this, the hierarchical CAR Bayes models

are used to explicitly address spatial correlation. The modeling results are given for Alexandria, VA, comparing modeling over area with modeling over roads. I use these models to predict crime/assaults and compare with the observed data.

Random forests (conditional and unconditional) are introduced in Chapter 6 in the context of using their variable importance measures to discover which covariates in the model are the strongest predictor variables. Conditional random forest variable importance measures are calculated for Alexandria, VA and San Francisco, CA. Principal components analysis is also used to create linearly independent combinations of the variables most highly correlated with crime. Finally some examples of multivariate visualization using DPnet with Alexandria, VA are given. I will give my conclusions, general challenges, and future work in Chapter 7.

CHAPTER 2. BACKGROUND

2.1 Crime Analysis on "Micro" Units of Place

Before proceeding with the research on analysis of crime data, it is helpful to be aware of the criminological theories that motivate analyzing crime using this local unit of analysis. Historically, most crime literature focused on large areas of the map for summary and represents crime as points. Recent criminology theory motivates looking at more "micro" levels, or small units of geography. This, in part, motivates the approach of using street segments as the fundamental geospatial unit of analysis.

Much early crime literature focused on person-oriented criminal propensity and crime at community levels such as states and neighborhood (Nettler, 1978). The spatial analysis of crime is a concept in part developed by Cohen and Felson (1979). In this paper they develop the routine activities theory, arguing that crime rates are not simply affected by the number of motivated offenders, suitable targets, and absence of security measures, but also by how often these three things come together in space and time. Sherman, Gartin, and Buerger (1989) use this theory as motivation for what they call the **"criminology of place"**, zeroing in on the analysis of the places where crime occurs. In their study, they found that about half of all calls to the police in Minneapolis, Minnesota came from only 3.3% of all addresses, with 4,166 robbery calls coming from just 2.2% of addresses. In another study by Weisburd et al. (2004) in Seattle, Washington, 50% of

crime incidents reported came from under 5% of street segments. These studies show why approaching crime at smaller units is important, as the interactions of offenders, targets, and security measures often occur in very specific geographic areas. Looking at larger geographic/political areal units (i.e. zip codes, census tracts) could mask a hotspot resulting from a few specific street segments within those units.

Weisburd et al. (2012) conclude 5 main points in their book about the criminology of place:

1. Crime is tightly concentrated at "crime hot spots", suggesting that we can identify and deal with a large proportion of crime problems by focusing on a very small number of places.
2. These crime hot spots evidence very strong stability over time, and thus present a particularly promising focus for crime prevention efforts.
3. Crime at places evidences strong variability at micro levels of geography, suggesting that an exclusive focus on larger geographic units, like communities or neighborhoods, will lead to a loss of important information about crime and the inefficient focus of crime prevention resources.
4. It is not only crime that varies across very small units of geography, but also the social and contextual characteristics of places. The criminology of place in this context identifies and emphasizes the importance of micro units of geography as social systems relevant to the crime problem.

5. Crime at place is very predictable, and therefore it is possible to not only understand why crime is concentrated at place, but also to develop effective crime prevention strategies to ameliorate crime problems at places.

Brantingham and Brantingham (1995) further describe the criminology of place by defining the various types of places in which crime can occur. A location that attracts criminal activity because of its social and physical geography, regardless of the level of criminal motivation that an offender may have, is known as a **crime generator**. There are many different factors that make a specific location a crime generator, such as traffic, population density, and proximity to shopping areas and sporting events (Van Patten, McKeldin-Coner and Cox 2009 and Short et al. 2008). These places where large concentrations of people are drawn result in favorable settings for certain types of criminal acts. A **crime attractor** is a place where an offender is already aware of known criminal opportunities. Examples of this type of place include bar districts, prostitution areas and drug markets (Brantingham and Brantingham 1995). Unlike crime generators, criminal activity at crime attractors is often from motivated offenders coming from outside of the area. Crime-neutral areas neither generate crime by creating opportunity nor attract motivated offenders; instead they occasionally see crimes committed by locals of the area. Crime-neutral areas tend to be low-crime places, and are useful to compare with the other types of areas by identifying the important differences between them. All of these types of places help explain the advantages of looking at more local levels of crime.

Brantingham and Brantingham (1995) also define nodes and paths in their description of the criminology of place. Nodes are central places in a person's life, such as their house, school or place of business. People tend to commit crimes close to these nodes and the paths between them, especially robbers and burglars. These paths can include street networks and transit systems, which can strongly influence the distribution on crime. Beavon, Brantingham and Brantingham (1994) discuss how the scheme of street networks can influence the amount of certain types of crime. In their study, crime was found to be higher in areas that were more easily accessible and had more traffic, and lower in areas with the opposite situation.

Weisburd, Groff, and Yang (2012) measure crime in units of street segments in order to focus on the criminology of place at a "micro" level. They define street segments as occurring between two intersections and including both sides of the street. Street segments are advantageous units of analysis as they are easily recognizable units of space with well-defined boundaries. Other analyses that use large areal units for analysis (e.g. administrative units such as census tracts, blocks, and block groups) as opposed to linear features such as the street segment may hide the variability of crime rates among the streets within an areal unit. For example, if a census tract contains street segments with largely different crime patterns (in terms of crime counts), segment information is lost when aggregating to the census tract level, obscuring possible hot spot locations. Large areal unit also hide segment-specific crime-related variables, such as calls for service and foot and vehicular traffic rates along blocks, among others.

Groff, Weisburd, and Yang (2010) further explain how crimes vary across street segments in Seattle, Washington by calculating group-based crime trajectories over 16 years of data, following the work of Nagin (2005). This type of analysis clusters the street segments into groups according to distinctive features such as their crime rates (e.g. high, low) and the change in crime rate over time (e.g. increasing, decreasing, stable). They use Ripley's K to establish whether similar trajectories are found among neighboring road segments (i.e. if similar trajectories are clustered together), or if there is great variation from street-to-street.

Figure 1 shows an example from Groff, Weisburd, and Yang (2010) where there is a high variation in crime trajectory patterns between many of the neighboring road segments in downtown Seattle. They found that there is heterogeneity between neighboring street segments in a number of places, which provides even further motivation to examine crime at the street segment level rather than in larger units of area, and provides motivation for smoothing over road segments in visualization. Smoothing methods help to reveal patterns by borrowing strength from neighbors to reduce noise. The heterogeneity at some neighboring locations suggests looking for circumstances that motivate restricting the extent of spatial smooth along street segments.

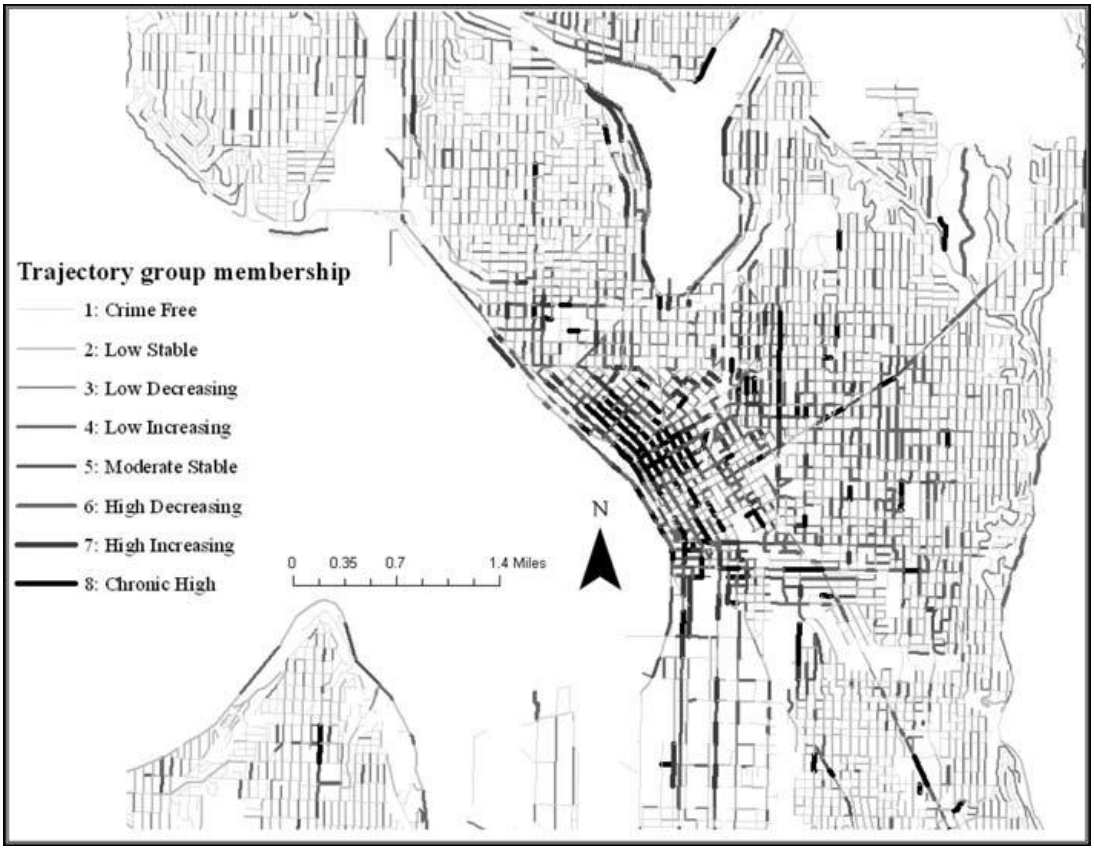

Figure 1: Spatial Distribution of Temporal Trajectories in Central Seattle (Weisburd, Groff Yang, 2010**)**

Further papers from these researchers support the approach of crime analysis at the micro level. Weisburd, Groff and Morris (2011) focus on hot spots of juvenile crime in Seattle. Their findings once again suggest that crime rates can vary greatly from one street segment to the next, and that targeting hot spots over street segments could help prevent crime. Another paper by Weisburd, Groff and Yang (2012) argues that focusing crime prevention on the level of specific street segments would be less costly and more effective than focusing efforts over larger areas.

In the Seattle study, Weisburd et al. (2012), analyze a number of covariates to see how they vary with crime. Some important variables included those that reflected crime opportunities across Seattle. These variables include the number of high-risk juveniles, residential population, the number of public facilities (e.g. hospitals, parks, etc.), bus stops, type of street (arterial v. non-arterial), police and fire stations, and percentage of vacant land. Other variables reflecting the social disorganization of Seattle include socioeconomic status, housing assistance, racial heterogeneity, and percentage of active voters, among others. This motivates the incorporation of such variables in my analysis of crime over street segments.

2.2 Previous Methods of Analysis and Hotspots

Hotspot mapping is a technique that helps to identify where the highest rates of crime occur. It is a predictive tool that uses past information in order to identify locations that need the most police patrolling. There are several different types of hotspot methodologies throughout the crime literature. An accepted standard has not yet emerged in this field. When visualizing crime data on a map, whether it is violent crime (e.g. assault) or damage to property (e.g. vandalism), certain neighborhoods appear relatively safe while other areas have dense clusters of criminal activity. Many criminals target the same areas repeatedly over a period of time. Crime "hotspot" analysis describes the areas that have larger than average crime counts or rates. With the increasing availability of data down to the incident level and the progress made in Geographical Information Systems (GIS), hotspot mapping has become a popular tool in

the crime analysis community.

The three basic types of visualization methods are frequently seen in crime literature are choropleth maps, spatial ellipses and smoothing. A choropleth map is a specific type of thematic map which refers to maps that have data aggregated to a political or administrative area, such as census tracts, zip codes or block groups. Crime events mapped as points can be aggregated within these geographic areas. An example of this can be seen in the paper by Anselin (1995) paper on local indictors of spatial association, which examines the spatial pattern of conflict in African countries (data is aggregated to country level). Another example in Eck (2005) shows vehicle crime data thematically mapped by census tracts (See Figure 2). Using crime counts is not adequate here, as a certain census tract may only have a higher count of vehicle crimes because the tract covers a larger population. While Census tracts were designed to have an average of about 4,000 inhabitants, they can vary between 1,200 and 8,000 people. Choropleth maps can be useful in representing certain summary statistics for political and administrative regions. However, the variation of spatial attributes within the regions (e.g. differing population sizes of census tracts) can make them inconsistent in portraying the underlying spatial structure without further statistical adjustments.

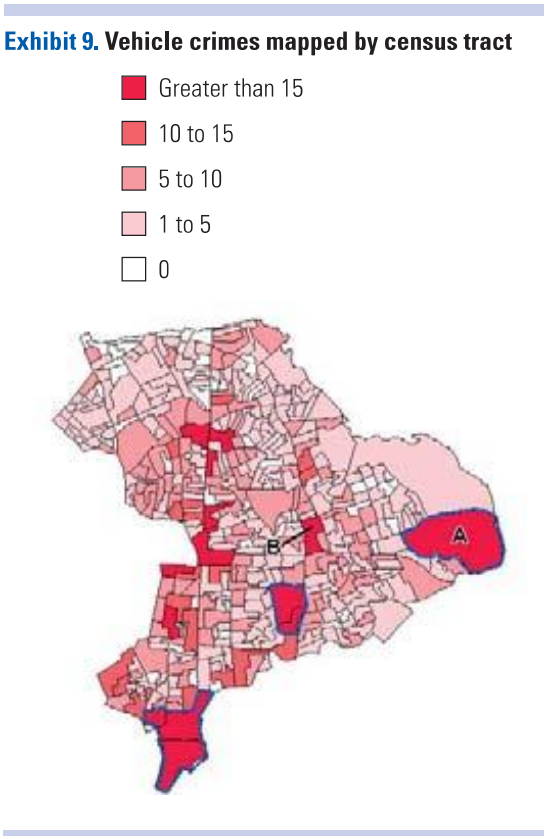

Figure 2: Thematic map of vehicle crime by census tract (Eck, 2005).

An example of using spatial ellipses involves nearest neighbor hierarchical clustering; this creates hot spot ellipses based on the nearest neighbor points that are closer than they would be if they had occurred by chance (complete spatial randomness). Examples of this are included in Liu and Brown (2003) and Levine (2006). Figure 3 shows an example from Levine (2006) using nearest neighbor hierarchical clustering to create hotspot ellipses of crashes in Houston from 1999 to 2001 caused by driving while intoxicated. Within the hierarchy, first-order clusters can be described as hotspots constructed from nearest-neighbor individual incidents, while the second-order clusters

group the first-order clusters into larger hotspots. A weakness of the spatial ellipse method is that crime does not naturally form spatial ellipses; some areas in the hotspots may actually have low crime (Eck, 2005). Spatial ellipses also ignore some important spatial details such as large lakes that could be in the middle of the area. Crime is rarely occurring over a lake so it would not be accurate to have part of the ellipse fall over this space. Also, choosing different parameters can produce different results. This DWI crash example illustrates the dominance of thinking in terms of points and areas in cartographic contexts. It would seem that a more focused map would show crash hotspots along road segments since they are primarily on roads and intersections rather than the elliptical regions.

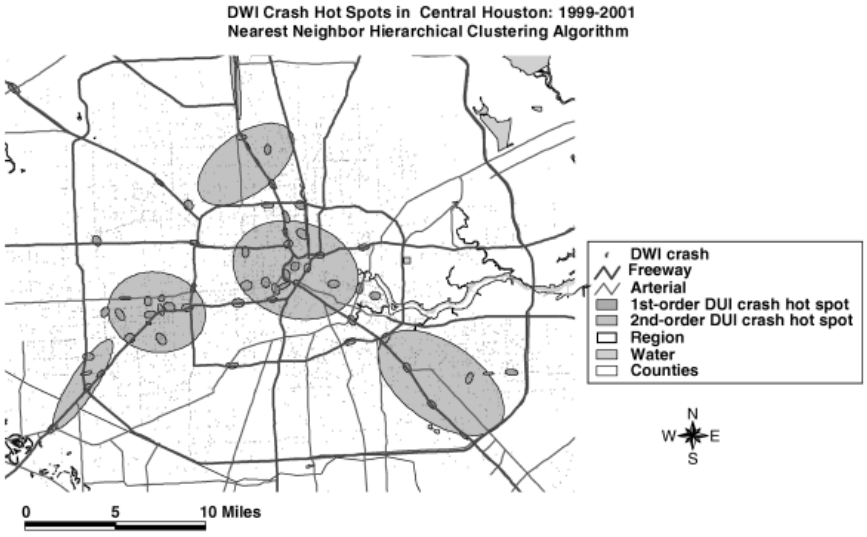

Figure 3: Ellipses from nearest-neighbor hierarchical clustering (Levine, 2006).

2.3 Smoothing methods and tools

2.3.1 Smoothing over areas

Smoothing helps to reveal patterns in data by approximating a function over the data that reduces the noise that distracts from the overall pattern. This flexible method tries to improve the precision of the area data without introducing large level of bias (Haining, 2003). Basic smoothing methods include local moving mean/median smoothers. For example, the moving median smoother replaces the value for case i, $z(i)$, with the median value from a set of values within a certain window of i, including $z(i)$ (Haining 2003). Similarly, the moving mean method replaces each point with the mean of a certain number of adjacent points; the larger the number of adjacent points included, the smoother the result. Oversmoothing introduces bias by reducing the local variation that is not noise. In mapping, smoothing creates a smooth surface $z = f(x, y)$ that can reflect counts/occurrences (in my case, crime data) or continuous data. Variable transformations such as counts per unit area support making comparisons. These can be represented in colored choropleth maps and in perspective surfaces.

Common graphical smoothing often uses weighted averages. The simplest spatial weighting considered is a simple binary weight:

$$w_{ij} = \begin{cases} 1, & if\ sites\ i\ and\ j\ are\ connected \\ 0, & if\ sites\ i\ and\ j\ are\ not\ connected. \end{cases} \quad (1)$$

Sites that are connected are considered spatial neighbors; what it means to be connected is open to interpretation. For example, counties can be considered connected if they

share a common border or if points within the county are less than a certain distance apart (Schabenberger and Gotway, 2005). This weight may instead be a function of other features, such as the length of the shared border, or population of the counties.

Distance weighting is another example of spatial weighting that uses the distance between points rather than the connectedness of areas. One distance weighting is the simple disk averaging, which assigns weights w_{ij} as follows:

$$w_{ij} = \begin{cases} 1, & if\ j\ is\ withing\ distance\ D\ of\ i \\ 0, & if\ j\ is\ not\ within\ distance\ D\ of\ i. \end{cases} \quad (2)$$

A distance decay function will give smaller weights to data values farther away and larger weights to those that are closer to the given data point. One example of this given in Haining (2003) is

$$w_{ij} = \begin{cases} \left[1.0 - \left(d_{i,j}/D\right)^{\beta}\right]^{\beta}, & if\ d_{i,j} < D \\ 0, & otherwise. \end{cases} \quad (3)$$

In this instance $d_{i,j}$ is the distance between cases i and j and β is a chosen constant. Weights are normalized to sum to one. All of these weighting methods can be incorporated into rules for smoothing maps. This notion of borrowing strength from neighbors appears in many statistical contexts, such as time series and bivariate kernel smoothing.

A smoothing method frequently used in the crime literature is kernel density estimation. Kernel density interpolation aggregates points within a certain radius and

then creates a continuous surface over the map to represent the distribution of crimes. Formally, if there are n independent observations $x_1, x_2, \ldots, x_n$ from random variable X, the kernel density estimator $\widehat{f_h}(x)$ used to approximate the density value $f(x)$ at point x is:

$$\widehat{f_h}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x_i - x}{h}\right), \tag{4}$$

where K is the "kernel" function and h is the bandwidth (i.e. neighborhood around point x) . Different types of kernel functions and bandwidths can be specified to give various resulting maps.

Eck (2005) uses quartic kernel density estimation, which incorporates the quartic kernel function, to create a smooth surface from the points over the map. Using a smoothing method such as this and creating a continuous surface makes it easier to interpret the general locations where crime is occurring. Levine (2006) gives an example of a three-dimensional kernel density interpolation of 1990 motor vehicle crashes relative to 1990 population in Honolulu (see Figure 4). The area-based smoothing across spatial locations used here is inconsistent with the phenomena because the values correspond to motor vehicles crashes, most of which occur on roads. If many roads are in the same area they "borrow strength" from each other but the individual roads, however the strength that they borrow is based on great arc distance over the entire space rather than distance along the roads. There could be many different barriers between roads where it would no longer make sense to borrow values from each other. Measuring area across bodies of

water, over a railroad track, or through the woods is not necessarily meaningful, as there would not be as much of a relationship between roads that face these barriers. For these reasons, borrowing strength based on distance along street segments would make much more sense for this motor vehicle crash data.

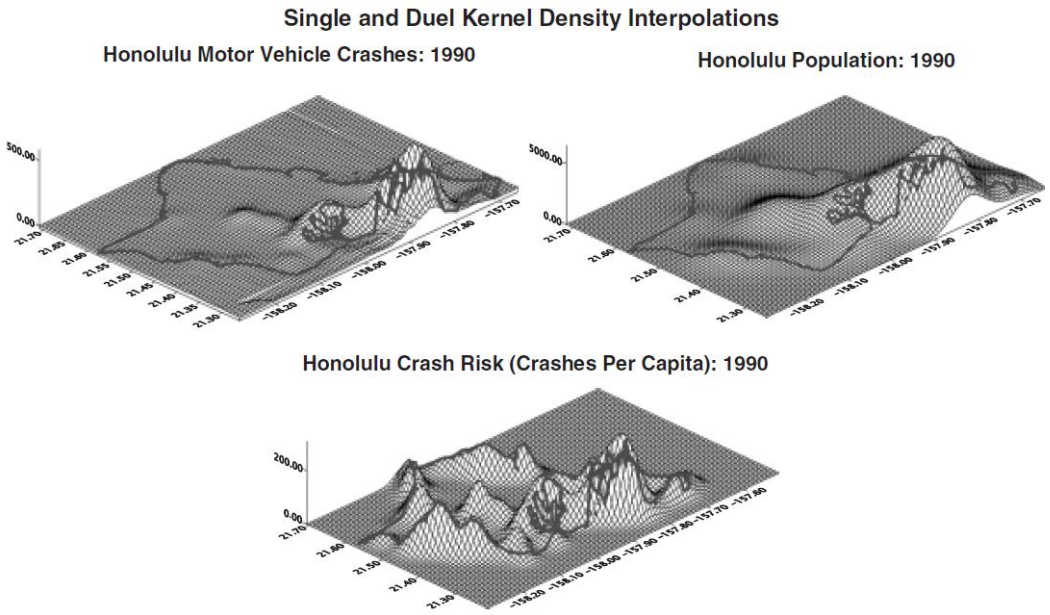

Figure 4: Kernel Density Interpolation (Levine, 2006).

Further crime smoothing methods include the transition density model described in Liu and Brown (2003). In this paper, Liu and Brown create a predictive algorithm using a point-pattern based density model, extending crime clustering methods by incorporating other variables based on criminal preferences derived from analysis of past events. Their transition density model measures the relationship between demographic, social, and

spatial attributes (among others) and measures of criminal activity, and uses a Gini-index-based measure for feature selection. The Gini-index is a measure of statistical dispersion of the data, with smaller index values indicating a higher level of cohesiveness and good set of features. In Lui and Brown (2003), this index I_g is calculated as follows. Define d_{ij} as the distance between event i and j in the feature subspace. Then a similarity score s_{ij} is calculated

$$s_{ij} = \frac{1}{1 + \alpha d_{ij}} \tag{5}$$

Where $\alpha = {}^{1}/_{\bar{d}}$ and $\bar{d}$ is the average event distance. Then the Gini index is given as variable this index is defined as:

$$g_{ij} = 4s_{ij}(1 - s_{ij}). \tag{6}$$

For a data set of n events, the averaged Gini index is:

$$I_g = \frac{2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} g_{ij}}{n(n-1)}. \tag{7}$$

Locations, times, and features of all incidents are the realization of a space-time point process. The transition density is divided into two components. The first order spatial transition density reflects the event intensity in the feature space, while second order spatial transition densities describe the interaction of a new event location with past event locations. Formally, let the features, location, and time of all incidents be a realization of

a space-time point process $\{x_{s,t}\epsilon\chi: s\epsilon D, t\epsilon T\}$, where $x_{s,t}$, s, and t are random quantities within feature space $\chi \subset \Re^p$, geographic space $D \subset \Re^2$, and time region $T \subset \Re^+$, respectively. Then calculate the density ψ_n by dividing the occurrence of events over time and space as

$$\psi_n(s_{n+1}, t_{n+1}|D_n, T_n, \chi_n) = \psi_n^{(1)}(s_{n+1}|D_n, \chi_n, T_n, t_{n+1}) * \psi_n^{(2)}(t_{n+1}|T_n) \qquad (8)$$

where $\psi_n^{(1)}(s_{n+1}|D_n, \chi_n, T_n, t_{n+1})$ is the spatial transition density and $\psi_n^{(2)}(t_{n+1}|T_n)$ is the temporal transition density. More information is given explicitly in Lui and Brown (2003).

Figure 5 gives an example of the resulting model, with darker shading representing areas of higher potential for crime created using Lui and Brown's density model. The white points represent breaking and entering crimes incident locations. This is a sophisticated model for handling spatial data with covariates; however, for my purposes, it does not take into account the distances over road networks that I seek to use. I did not pursue developing a special variant for roads.

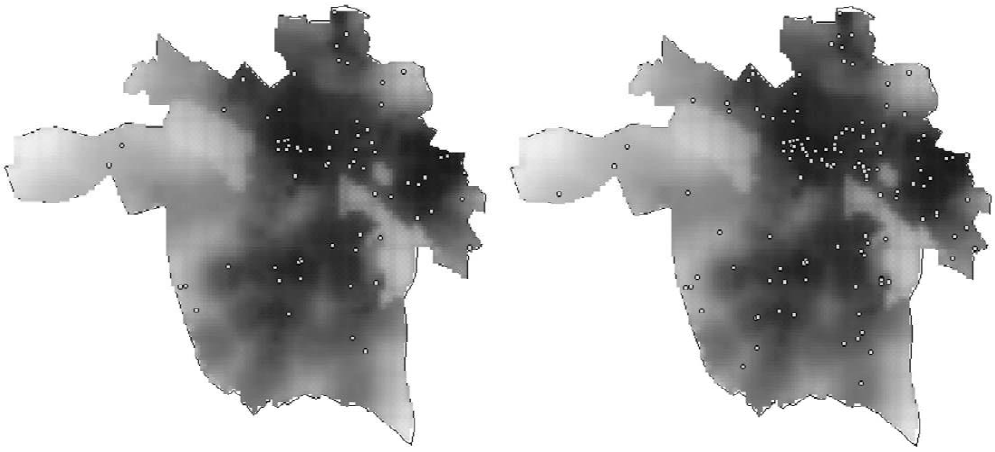

Figure 5: One of the resulting models calibrated on July 7–20 data, tested on July 21–27 data (left) and July 21–August 3 data (right) (Liu and Brown, 2003).

Smoothing methods are often used to call attention to regions with high and low values and can serve as hotspot detection algorithms. Some algorithms are specifically designed for hotspot detection. Here I call attention to the well-established spatial scan statistic method initially developed by Kulldorf (1997) and later extended in several ways. Spatial scan statistics uses a moving circular window on each centroid of a region, with the radius of the circle varying. If the window contains the centroid of a region, then that whole region is included in the window. These circles cover the map and may have many partially overlapped circles of different sizes. For each window, it is possible to compute the likelihood of observing the observed number of cases within and outside the window, respectively. Kulldorf (1997) proposes the spatial scan statistic for the Bernoulli and Poisson models. There are also many extensions to the spatial scan statistic that use ellipses and cylinders as opposed to circles.

2.3.2 Smoothing over networks

In mathematics, networks are a series of connected edges and vertices with a number or weight assigned to each edge. In real-world applications, networks can include streams, pipelines, streets, communication networks, and much more. Each type of network has properties that motivate the choice of analysis method refinements. For example, analysis of streams may incorporate weights based on flow direction. For analysis of street segment data, whether or not the traffic flow is one-way or two-way may make a difference depending on the phenomena being studied.

De Oliveira (2011) extends the spatial scan statistic over pipes in a water distribution network. He hypothesizes that the networks contain unexpected clustering of pipe breakage points, and attempts to discover where the clusters of breaks are located. While typically spatial scan statistics use aggregated count data, in this paper they focus on each individual break event, and define a cluster as a connected subgraph of pipe networks that has a significantly higher density of breaks than what is expected. Within the constrained space of pipe networks, the basic spatial scan statistics use a distance metric that will not accurately describe the space. De Oliveira creates another approach that relies on the shortest path distances between points along the pipe network.

Another paper that uses path length measurements is Curriero (2006), where stream distance is used as a foundation for kriging. Kriging is a method of interpolation where values at a point are predicted using a weighted average of known values in some neighborhood of that point. The stream-restricted and unrestricted network distances can be quite different. The length of the stream using stream distance was 134 miles, while

the length when not restricted to this path was only 70 miles. Results showed that kriging using the stream distance provided a more accurate prediction than kriging based on Euclidean distance.

Other papers that also look at spatial statistics along stream distance include Peterson, Theobald and Ver Hoef (2007) and Ver Hoef, Peterson and Theobald (2006). They make the claim that Euclidean distance may not be ecologically meaningful as this measure does not accurately represent the spatial configuration of a stream network. They develop a new measure using hydrologic distance, which is defined as the distance between two locations when movement is restricted to a stream network, and may or may not be limited to flow direction. Typical spatial autocovariance functions may not be valid when looking at non-Euclidean or arc distance, and these papers discuss ways of developing a valid model to surpass this obstacle.

Similar to these methods of smoothing over networks, I use nearest road segments as opposed to nearest areas. I use the road distances between midpoints of connected segments as weights in the spatial modeling.

2.3.3 Smoothing software tools

There are many software tools available to implement these hotspot and smoothing methods. For example, SaTScan is a free software that analyzes spatial, temporal and space-time data using the spatial, temporal, or space-time scan statistics (Kulldorf, 2009). SaTScan uses several different models, including both the Poisson and Bernoulli models along with the space-time permutation model, the ordinal model, and

the Exponential and Normal models. The software can handle data aggregated to census units or other geographical levels, or unique coordinates for each observation. SaTScan has the flexibility to handle spatial heterogeneity of a background population, any categorical covariates, multiple data sets, temporal trends and missing data.

CrimeStat is a popular spatial statistics program for the analysis of crime incident locations. CrimeStat was developed by Ned Levine & Associates of Houston, Texas and provides statistical tools to aid many law enforcement agencies and researchers across the country in effectively mapping crimes (Levine, 2013). The program includes more than 100 statistical routines for the spatial analysis of crime, including nearest neighbor hierarchical clustering, kernel density estimation, space- time analysis, journey-to-crime modeling and regression modeling (including Poisson regression), among other area-based methods, and also has the ability to write graphical objects to ArcGIS.

Ver Hoef, Peterson, Clifford and Shah (2014) created an R package that will analyze the stream networks described in their previous works. The Spatial Stream Network package (SSN) imports GIS data as a SpatialStreamNetwork object and uses distance metrics and geostatistical models unique to stream networks, including water volume and directional flow. The package also includes traditional models that use Euclidean distance, simple random effects models, and Poisson and binomial families for a generalized linear mixed model. The most unique component of the SpatialStreamNetwork object being analyzed is that it contains both point and line features within the same object rather than as two separate objects.

2.3.4 Visualization tools

The mapping techniques described above focus on visualizing one variable at a time. Conventional segment encoding does not extend to viewing multiple variables (e.g. crime rates and crime-related covariates) together in one visualization. There are examples that use segment thickness and dash patterns but these are not typically effective when the segments in the plot are dense. However, there is an effective dynamic Java package called CCmaps (conditioned choropleth maps) for viewing areas characterized a dependent variable and two covariates. Dr. Carr produced a variant called DPnet (dynamically partitioned network) that features colored polylines rather than color polygons. This is suitable for showing road segments. Carr, Wallin and Carr (2000) first described the conditioned map approach to representing three variables. Further descriptions with CCmaps examples appear in Carr, White and MacEachren (2005) and Carr and Pickle (2010). The basic idea when applied to road segments is to represent the dependent crime variable using color. There are three colors that distinguish low, middle, and high values. The analyst controls what is meant by low, middle, and high by using a three-class slider. To address two covariates, DPnet partitions the single map into a two-way 3 x 3 grid of maps. The grid highlights road segments with low, middle, and high values of the first covariate in the left, middle and right columns respectively. Analogously, the grid highlights the road segments with low, middle and high values of the second covariate in the bottom, middle and top rows, respectively. Technically, showing the non-highlighted segments in a 4th color that is

closer to the background serves to highlight the segments shown in the color slider colors.

The analyst selects the dependent variable to attach to the color slider and the two covariates to attach to the two three-class partitioning slides. Then the analyst adjusts the sliders to control what is meant by low, middle, and high values. Dynamic feedback includes changes in road segment colors in the grid of maps. The averages of the highlighted road segment crime values for each the nine map appears at the top right of each of the map and the R-squared from fitting these means to corresponding partitioned sets of road segment crime values appears at the lower right of the grid of maps. This and other statistics are all updated dynamically with adjustments to the partition slider thresholds. Examples appear in Chapter 6 using DPnet to partition crime counts for road segments in conjunction with two additional variables.

CHAPTER 3. DATA

3.1 Crime Data

I use and develop methodologies that support multivariate analysis and visualization of crime statistics with road segments as the fundamental unit of analysis. However, such methodology is of little use without associated data. The collection, origination, and access to data are crucial to model. Data availability and quality has a major impact on the resulting models.

The crime data I use in illustrating this methodology comes from the Alexandria Police Department for Alexandria, VA. The City and County of San Francisco also provided data for San Francisco, CA that can be accessed directly on the web (data.sfgov.org). The methodology includes converting local geospatial polygon data such as U.S. Census block statistics (U.S. Census Bureau, 2013) to polyline (road segment) data. Having additional variables associated with road segments opens the door to multivariate parametric and nonparametric modeling and prediction.

3.1.1 Alexandria Crime Data

The Alexandria Police Department provided a data set of crimes reported for the years 2006-2010, with examples of some of the variables given in Table 1. Variables in this data set include a description of the crime committed, classification of that crime, and

the time of day and date which the crime occurred. The time reported is generally recorded within the hour after the incident concludes; however, sometimes the report time or date is several days after the incident (for example, in cases of rape where there may be a delay in the victim reporting). Time of day can factor into specific types of crime, such as robberies around midnight or daytime residential burglaries. Crime is broken down into Part One, Nuisance, and All Other Offenses. Within those categories there are subgroups of Part One (murder, rape, robbery, felonious assault, burglary, larceny, motor vehicle theft, etc.), Nuisance (alcohol violations, drug, prostitution, destruction, gambling, disorderly, etc.), and All Other Offenses, which includes anything that doesn't fall into the first two categories.

Table 1: Description of Alexandria Police Department Data

Variable	**Description**
INCINMBR	Incident Number
DTREPORT	Date of Crime (From 1/1/06 to 12/31/10)
TMREPORT	Time of Report
INCLASLIT, OFFENSELIT	Type of Crime: Assault and Battery, Residential Burglary, Driving While Intoxicated, etc.
AL_CRIMES_DB_CRIMES_IBRGENERAL	Grouped Crime Types: Assault Offenses, Larceny/Theft Offenses, Liquor Law Violations, etc.
CENSUS, SUBCEN	Census ID numbers
LAT, LONGIT	Northern Virginia State Plane Coordinates (in feet)
ZIP	Zip Code 22314, 22301, etc.
LOCATION	Address of where each crime occurred/ was recorded
AL_CRIMES_DB_OFFENSES_IBRAGAINST, AL_CRIMES_DB_OFFENSES_IBRGENERAL	Crime against Person, Property, or Society

Variable	Description
LOCTYPELIT	Type of Locality: Highway/Road/Alley, Residence/Home, Parking

There are a total of over 62,000 crimes recorded, with both addresses and *x-y* coordinates given for where the crime occurred. The coordinate system used here is the State Plane Northern VA coordinate system, which gives values in feet. This coordinate system is the preferred projection of typical geospatial coordinates (latitude and longitude) chosen to most accurately preserve distance between points with minimal distortion to areas and angles. Crime points that are not directly on top of the street segment are geocoded to the address where the crime occurred. Estimated points, (for example, 400 King St.) are geocoded to sit on the actual street. These are approximate points where the crime occurred; under certain circumstances an exact location may be unknown. A few addresses with crime counts were removed because it was known that those crimes did not occur at these locations. These include 2001 and 2003 Mill Rd and 2034 Eisenhower Ave, which are locations of the old Police Department/Sheriff's Offices. Missing data exists where there where geocoding errors; some points were geocoded to be outside of the city limits of Alexandria and were removed.

I will analyze both the full crime data set and assault offenses. Assault offenses include assault and battery, simple assault, felonious assault, and assault and battery of a police officer(s). This focuses on crimes involving the threat and/or an occurrence of bodily harm, as opposed to crime overall that includes both violent and nonviolent

crimes. Assaults are also relevant over road segments as many assaults occurring over roadways (road rage assaults) and outside of public facilities near the street. This subset includes over 5,500 crimes.

One tool that allows us to visualize the crime location on a map is the R package '**RGoogleMaps'**. Figure 6 shows a subset of the City of Alexandria crime data with the crime locations plotted using yellow triangles.

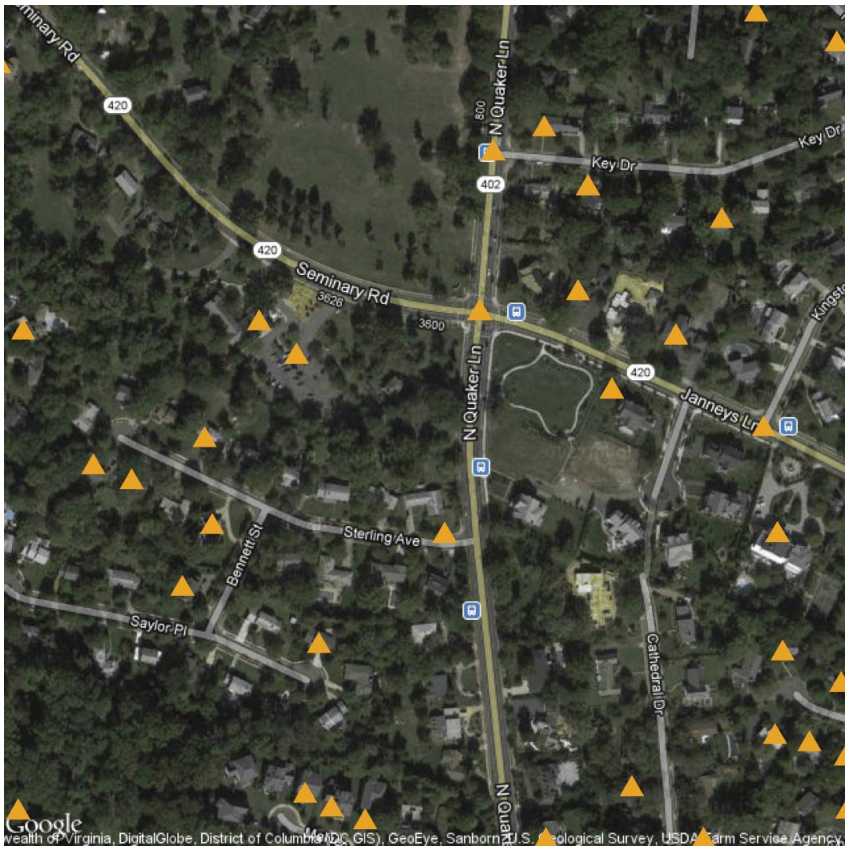

Figure 6: Close-up of Alexandria Crime using RGoogleMaps.

I also show the Alexandria crime data using ArcMap from ArcGIS (ESRI, 2011). Figure 7 gives one example using Assault Offenses. There are at least three sections of the city in which distinct clusters appear. One section is Old Town Alexandria, which is on the Southeast side of the city by the Potomac River on the right side of the map. This is the downtown area of the city with many streets and businesses close together. Next is West Alexandria, which is on the left-hand side of the map. With this particular map view, the West Alexandria cluster could be divided up into further clusters, for example surrounding the highway on the left and one more towards Central Alexandria. The third section is North Alexandria, which is technically in the Northeast section of the city.

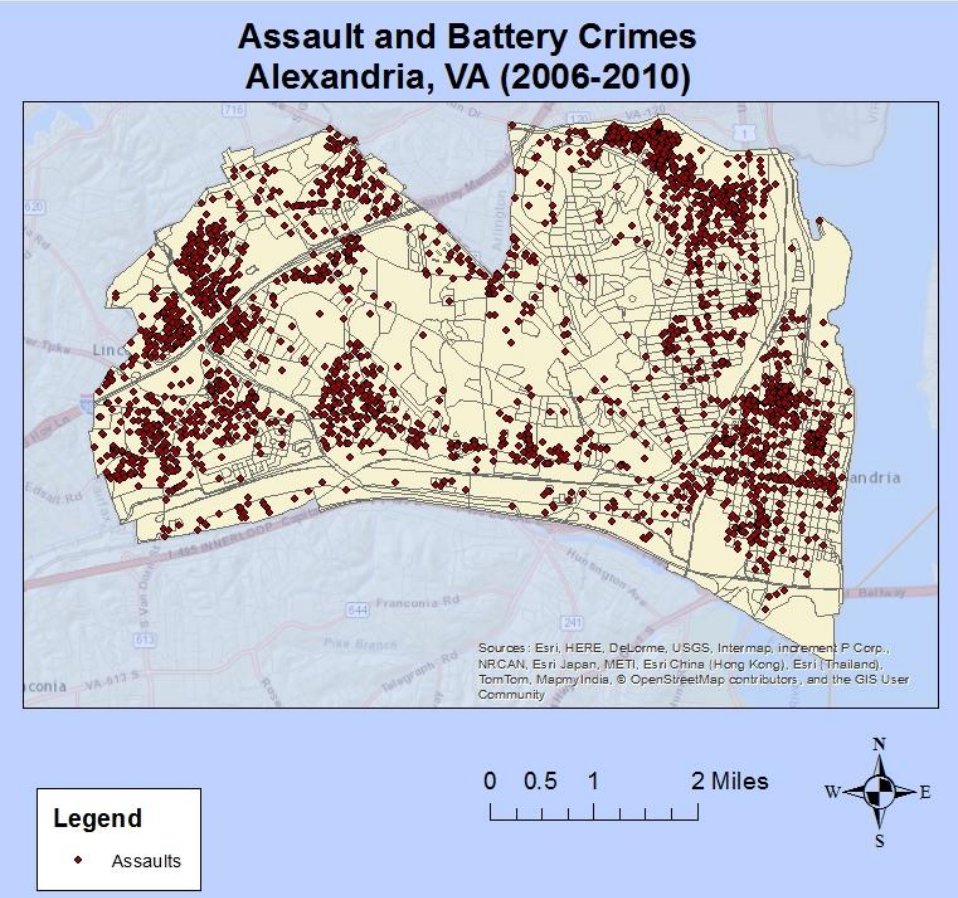

Figure 7: Assault and Battery crime in Alexandria as viewed in ArcMap.

Note that this visualization does not take into account the number of overlapping points; some points actually represent a location where many crimes occurred. I address this by associating these crimes with the nearest road segments and assign the road segments a color scale based on the number of crimes along each segment. Figure 8 is a map of assault counts along each road segment divided by the length of that segment in miles. This supports thinking in terms of crimes per mile and helps to adjust for the tendency of longer segments to have more crimes. Colors are defined by the quintiles of

this value. Blue road segments correspond to low crime density while red road segments correspond to high crime density. The red road segments clearly corresponding to the three clusters mentioned previously (that is, Old Town, West, and North Alexandria). The map legend shows the upper bound on road segments is 511.30 assaults per mile. Such segments beg for explanation. Are there many counts, a small length, or both? Are there location reporting errors? We would not notice this anomaly in Figure 7. In terms of graphical representation of point density, the direct plotting of points has poor perceptual accuracy of extraction (Cleveland and McGill, 1984). It is better to use methods that show crime density along road segments as seen in Figure 8.

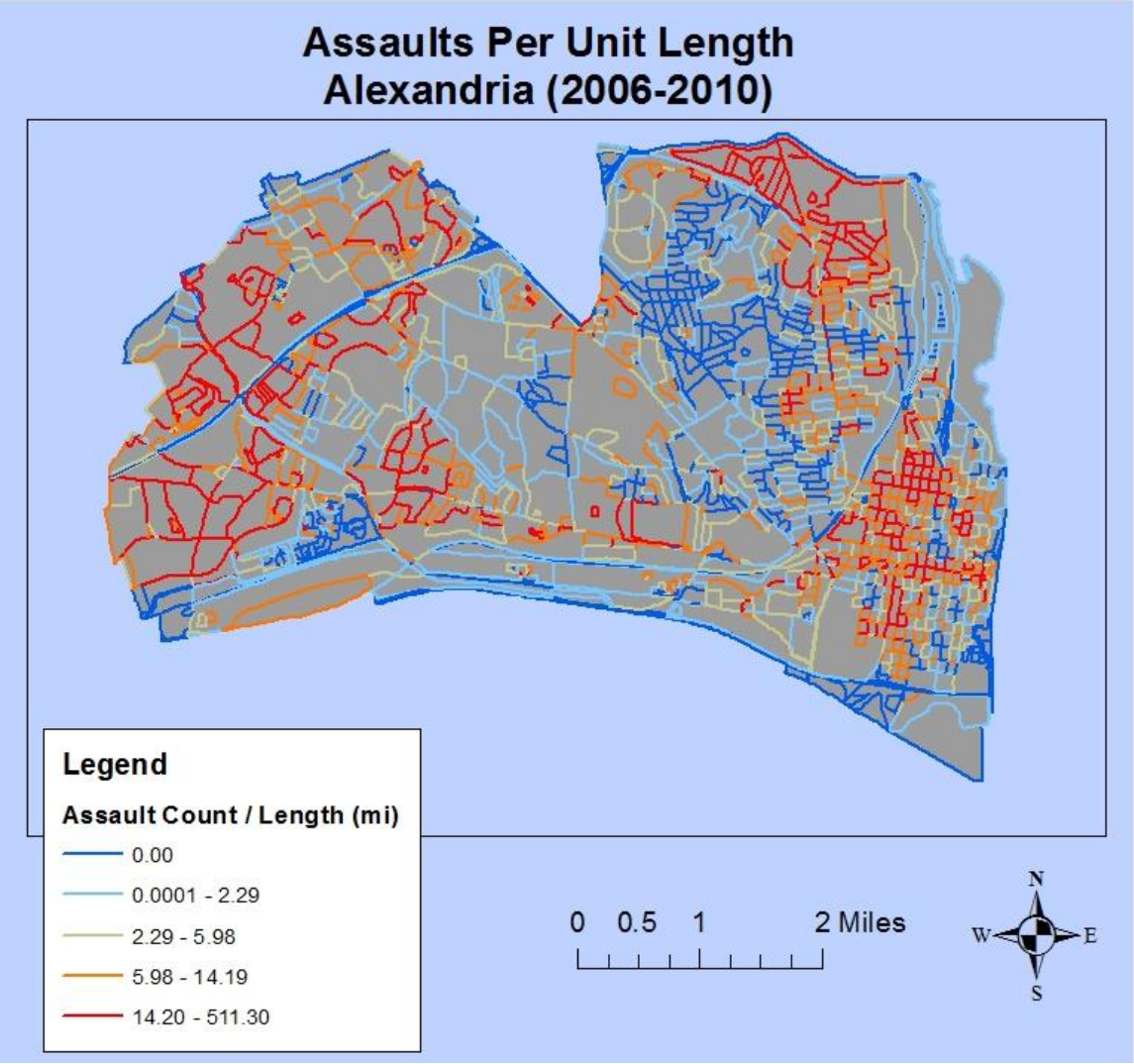

Figure 8: Alexandria Assault and Battery Crimes with crime count represented by road segment color.

3.1.2 San Francisco Crime Data

Crime data locations for San Francisco, CA are made freely available by the City and County of San Francisco at data.sfgov.org (2014). Crime data sets for a big city such as San Francisco are much larger than for Alexandria, VA. I downloaded the crime incidents for 2012, which includes over 123,000 crime points. Variables in the San

Francisco crime data set are given in Table 2. Crime points are not as accurately geocoded as in the Alexandria data set; that is, they are only geocoded to the nearest block (e.g. 900 Block of Hyde Street) or intersection (e.g. 6th St/ Harrison St).

Table 2: Description of San Francisco Crime Data

Variable	**Description**
IncidntNum	Incident Number
Date, DayOfWeek	Date and Day of Week Crime was Committed (From 1/1/12 to 12/31/12)
Time	Time of Report
Descript	Description of type of crime
Category	Grouped Crime Types: Vehicle Theft, Assault, Robbery, etc.
PdDistrict	Police District: Central, Southerm, Bayview, etc.
Resolution	Was someone arrested, cited, etc.
Location	Block where each crime occurred/ was recorded

Figure 9 shows a map of all of the crime points (in red) in San Francisco. The northeast section of San Francisco is the downtown area, which is down the bottom of a large hill. Based on past data, larger counts of crime are expected in this area. In Figure 9 there is such a multitude of points over a small area that they completely cover the map and make it impossible to visually assess crime density.

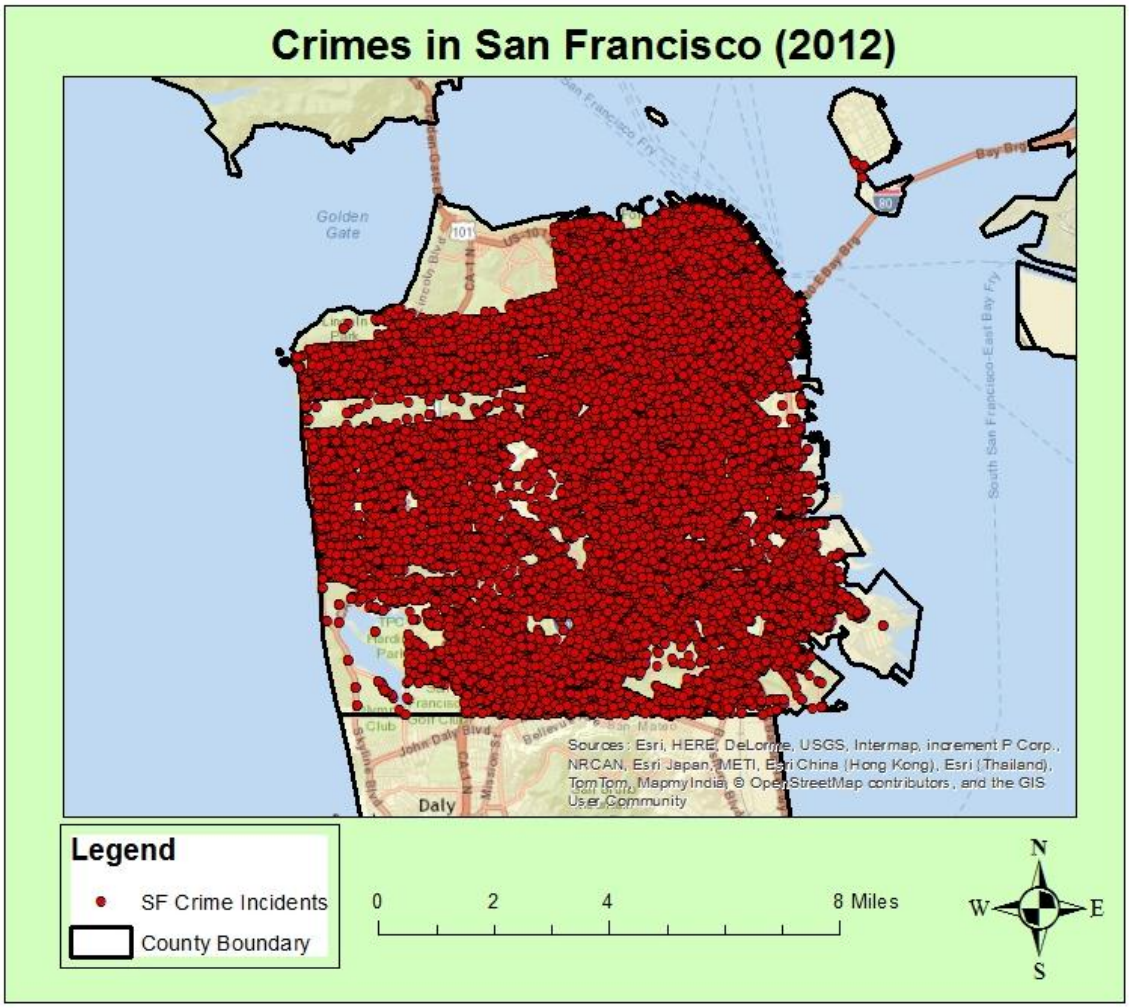

Figure 9: Crimes in San Francisco, CA in 2012.

I can once again map the crimes to the roads to get a simple visualization of where most of the crimes are located. Figure 10 displays the crime by length similar to the map in Figure 8. It is now possible to see the higher crimes located in the downtown area of San Francisco by the red road segments.

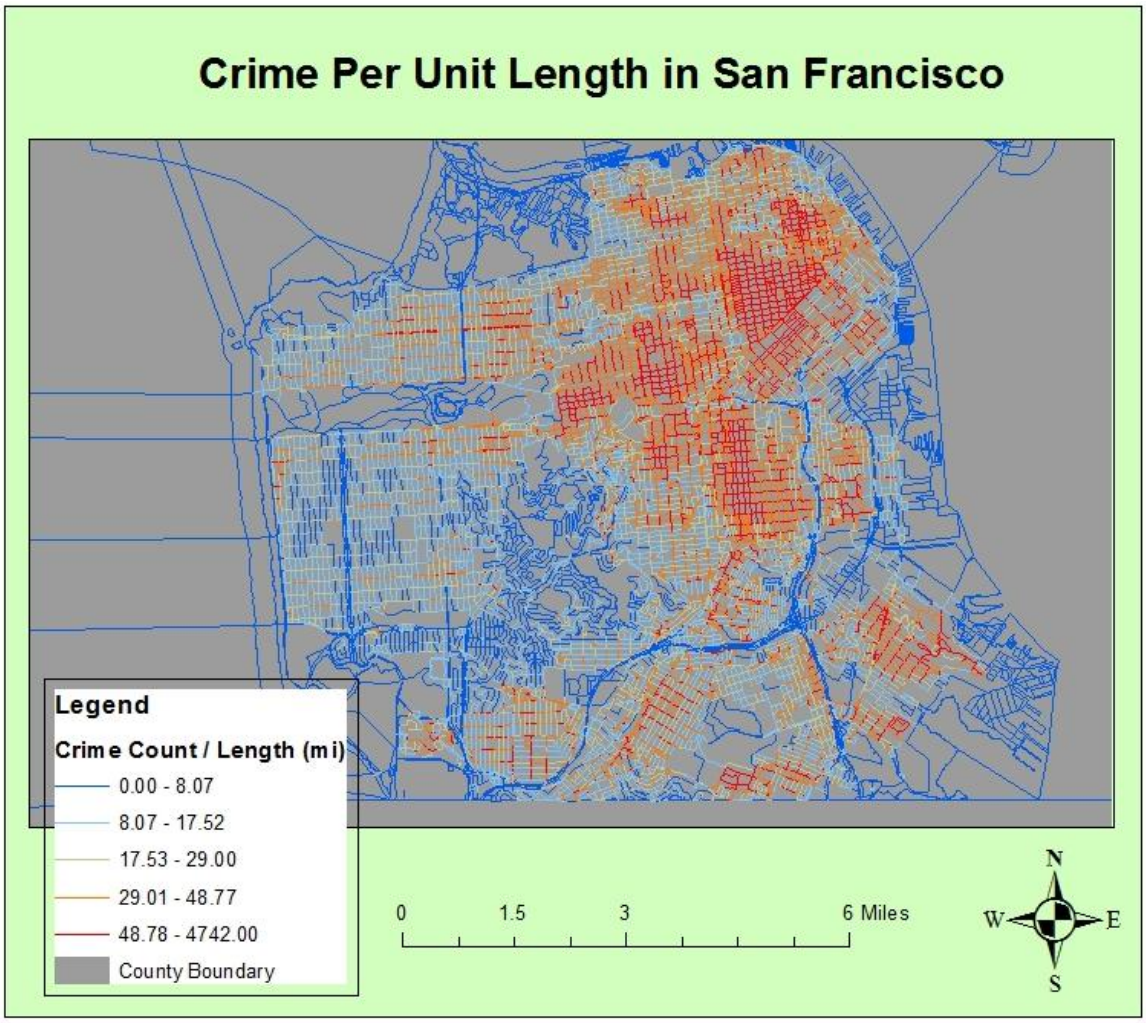

Figure 10: San Francisco Crimes with crime count per unit length represented by road segment color.

3.2 Crime-Related Variables

Data availability and quality limit the variety and quality of the models that help reveal and explain patterns in the data. The criminology of place emphasizes the value of micro analysis, so motivates seeking point data with specific locations, road segment data, and area data at the smallest areal units available, blocks. A rich resource for block data is the U.S. decennial Census. This is the most spatially detailed data set the Census

provides. There are a total of 1,294 blocks in Alexandria. There are some locations where rapid change of these area statistics could be indicative of social disruption that relate to change in crime rates at the segment level. A limitation of the block data is that it spans 10 years. There are some time mismatches when comparing crime (e.g. years 2006-2010) with data from the decennial census (2010 based on 10-year time span). However, the stability of the block data statistics and crime at locations over time still makes this block data relevant. Statistics on age, sex, residential population and housing units can be found on the U.S. Census FactFinder website for 2010 at the block level (U.S. Census Bureau, 2013). I downloaded shapefiles for Alexandria down to the block level from the 2010 US Census to map this information, and this downloaded data I transform to variables used.

Age is a useful variable since most crimes are committed by those of a certain age. Across age groups, there is a rise in the number of crimes committed by offenders in their adolescent years, with a peak in offenses by those in their late teens/early twenties, and a steady decline after that (Vold, Bernard and Snipes, 2002). The U.S. Decennial Census provides statistics for approximately 20 age categories that I collapsed to the following 5 basic categories: Under 17 years old, 18-24 years old, 25-44 years old, 45-64 years old, and over 65 years old. I also produced an additional variable by converting the counts to the percent of each age group within each block as an additional variable to consider in modeling. Figure 11 gives an example of the 18-24 year old age group in Alexandria, VA. The units here are percents converted to decimal form. Notice how the

areas with high percentages of 18-24 year old are similar to the crime clusters shown before.

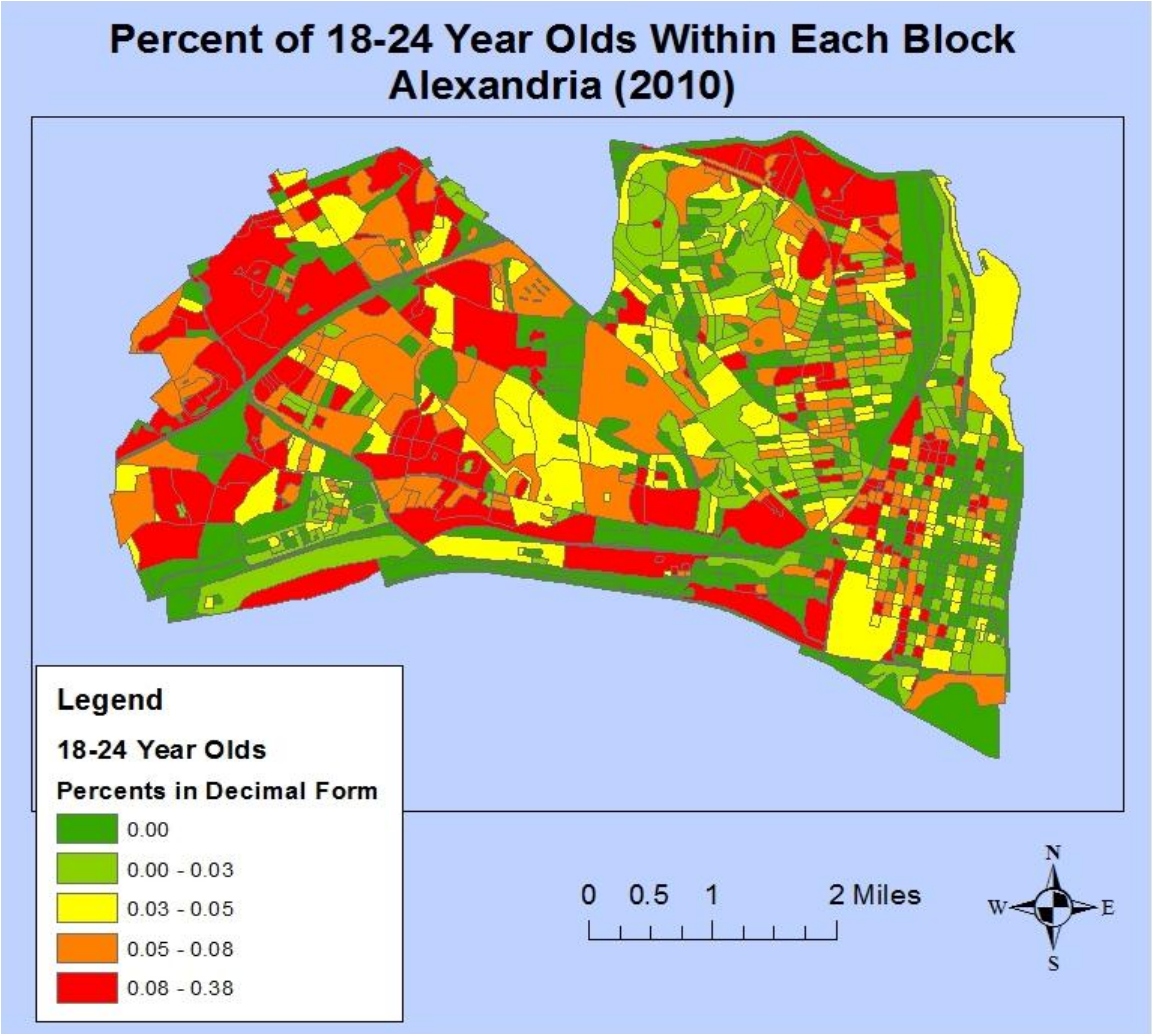

Figure 11: Age within each block in Alexandria, VA.

The Census also provides gender. I convert the count data within each block to percent male and percent female as an additional variable to consider in modeling as

well. Gender has one of the highest correlations with crime, with males much more likely to offend than females (Vold, Bernard and Snipes, 2002). I convert the population counts to population density by dividing the counts by their calculated block areas. In crime literature, the large population of certain neighborhoods is found to be associated with high crime rates (Vold, Bernard and Snipes, 2002). Finally, I looked at housing block data and converted housing units to housing density by dividing by the area of each block. Population and housing densities help describe the overall environment, which are important factors in environmental criminology (Brantingham and Brantingham, 1995). I use ArcGIS to assign the area statistics to each polyline surrounding that block. Each road segment will have two values. I will use the average of those two values for modeling and visualizations.

3.2.1 Alexandria Crime-Related Variables

Aside from the Census block data, I obtained a few other crime-related variables for Alexandria, VA. Two members of the Alexandria Police Department provided calls for service data. This data is for the same time frame as the crime data (2006-2010) and included over 50,000 observations in an Excel file. I successfully geocode about 39,000 of these call locations using ArcMap in ArcGIS. The location given only includes nearest intersection where the call was made rather than the actual address. I used ArcMap tools to assign each call to the nearest road segment, with each segment getting a sum of those calls. This data set gives the date and time of the call and the type of call. The type of call indicates witnessing something suspicious, a noise complaint, disorderly

conduct, shoplifting, etc. I expect this variable to by highly correlated with crime data, as many of the arrests made are a result of a call made to the police department. When analyzing assault crimes, I subset the calls for service data to only include calls related to assault. For modeling purposes, I also created a variable by removing the "crime-related" calls but keeping complaints and other descriptors of social disorder. More specifically, this "social disorder" variable includes animal complaints, drug complaints, missing person reports, noise violations, parking complaints, suspicious events, suspicious packages/substances, telephone complaints and traffic complaints. An article from Wilson and Kelling (1982) makes the claim that disorder and crime are strongly linked at the community level. In this article they describe what is known as the "broken windows" theory; that is, if a window is broken on a building and not repaired, the rest of the windows will soon be broken. Undesirable behaviors in the neighborhood, if left untended, will lead to a breakdown of community control and lead the area to become more vulnerable to criminal activity. I associate this variable with the nearest segment to give another resulting plot of count per unit length, shown in Figure 12.

Home prices in Alexandria were obtained from GMU Center for Regional Analysis. (Dr. Ed Zolnik and Jeanette Chapman) through the Metropolitan Regional Information Systems (MRIS). This data include homes sold from 2006-2010, which includes over 7,000 homes in the Alexandria area, geocoded to the exact address of the home. It must be kept in mind before analyzing this data set that 2006 was the peak of the housing bubble, so home prices will be higher in 2006 and then decrease/flatten out over 2007-2010. The number of sales will also be higher in 2006 and then drop off. This

data set includes townhouses, duplexes, single detached homes, and condos, but excludes most new homes and all rental units. There is a lot of research on the relationship between crime and economic conditions/poverty (Vold, Bernard and Snipes, 2002) and the price of homes can be a good reflection of this relationship. This variable, since it is in point data form, can also be associated with the nearest segment. The number of vacant houses is a useful variable in other studies, but I was not able to obtain this.

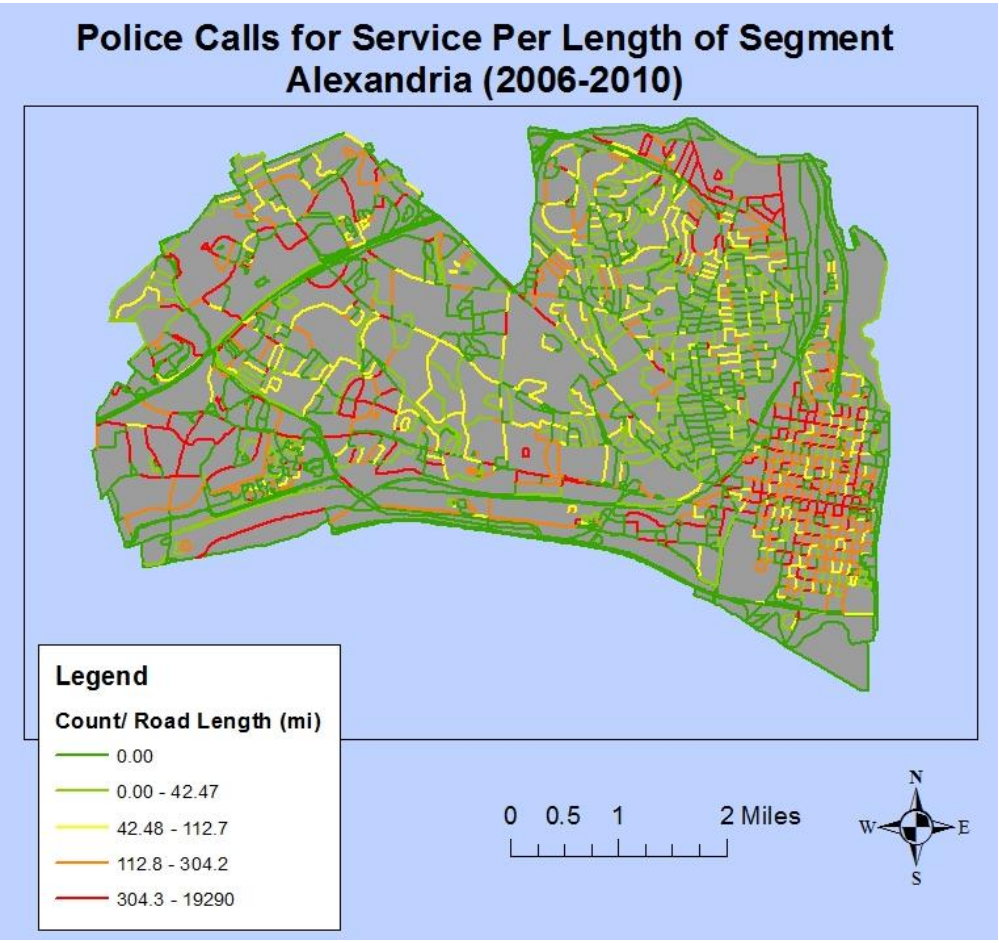

Figure 12: Police Service Calls by Length in Alexandria, VA.

I studied the data to pick variables that were needed and transformations to improve the modeling. After assigning the crimes to the road, I round the fractional values to the nearest integer (CrimeInt and AssaultInt) because the Poisson and Negative Binomial models will only work properly if given counts as the response variables. Out of the different age and gender categories, I specifically select the counts of 0-17 year olds, 18-24 year olds and the count of males for use in the model. I select these age and gender variables out of the possible categories since there is strong background literature of the influence of young males and crime. I will also use an interaction term specifically with the 18-24 year olds and males to account for the strong relationship the two variables jointly have with regards to crime.

The population and housing densities are scaled by dividing by 1,000 (since the area in meters squared created very small values). The full calls for service data set are used for the full crime data set, while the assault calls for service data is used for the assault data set. The social disorder variable is used for both data sets. The distributions of these variables are highly skewed. To reduce this skewness in the data I use a simple square root transformation of the age, gender, housing, population, and call count variables. Housing prices are only available for selected segments. We will use a simple imputation strategy to give values to segments that do not already have values. Imputation is the process of replacing missing data with substituted values. For those segments that do not have housing prices, I give them the average value of the prices from two levels of nearest-neighbor segments. If none such value exists at this level, I give the segment the median housing price value of the entire data set. More

sophisticated imputation methods were not used here, but are possible. These housing price values are also scaled by 1,000.

I create a scatterplot matrix of the crimes versus the crime-related variables using hexagon binning. In hexagon binning, the entire xy plane is divided into a grid of hexagons, with the number of points falling in each hexagon being counted and stored. The color in the hexagon plot is determined by this count, with higher count densities getting a darker color. The matrix also does a loess smooth over each plot, which gives a smooth line to represent the overall trend in the data. The diagonal plots show the variables names and their densities. The bottom row shows how the covariates relate directly to the crime variable at the left of the row. You can see, for the most part, a very small increase in values as crime increases.

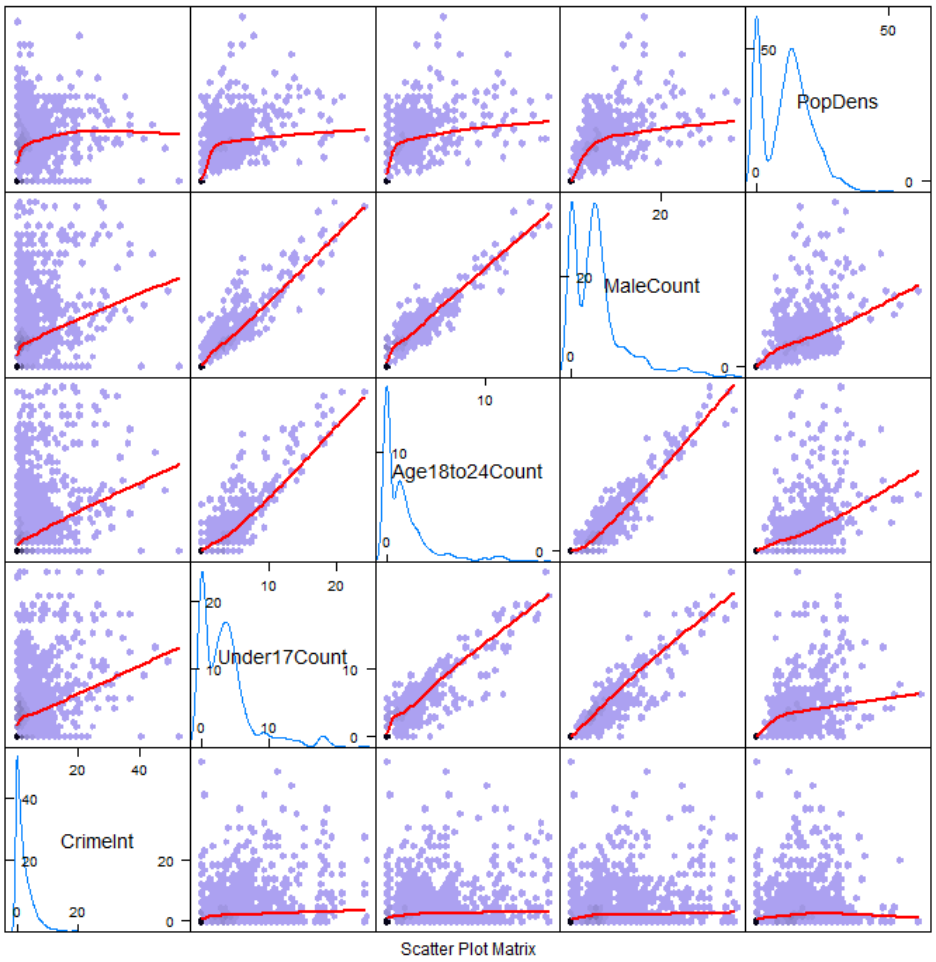

Figure 13: Scatterplot matrix with hexagon binning and loess smooth: Full Alexandria crime data set, first subset of variables.

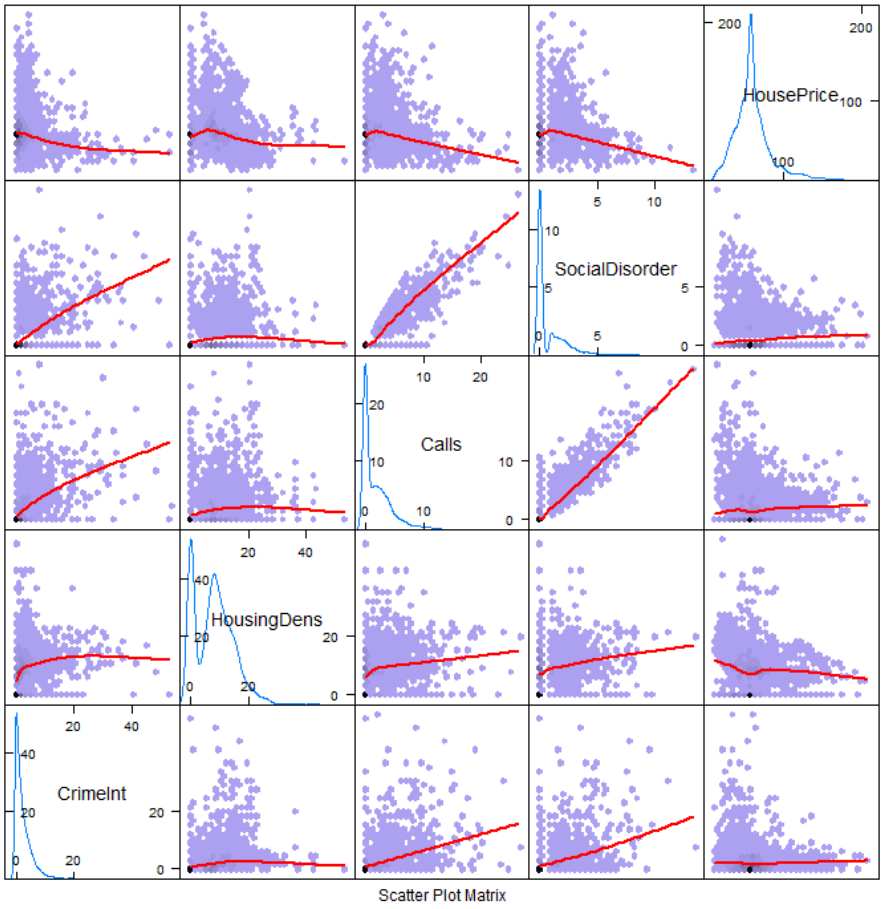

Figure 14: Scatterplot matrix with hexagon binning and loess smooth: Full Alexandria crime data set, second subset of variables.

Each variable has a positive correlation with crime except for housing price. The crime decreases in areas with high property values. Crime increases as population and housing density increases but then at a certain point starts to stabilize. There are many zero values for many variables. The high number of zeros leads to a dark hexagon at the lower left of each plot. This supports the use of zero-inflated models, described in Chapter 5.

Figure 15 highlights an outlier with a green dot. This road segment has a large number of calls for service compared with its number of crimes, going against the trend of the other segments. I extract the geographical coordinates of this segment and identify it using RGoogleMaps. This segment is near a group of large apartment buildings along the highway, and nearby the Landmark Mall (opposite side of the highway on the southeast corner of the map), which tends to have large counts of crimes located near it. Identifying individual anomalies in this way really gets the root of the criminology of place; each segment can be very unique and different. We do no delete this outlying observation prior to modeling, however this can be taken under consideration.

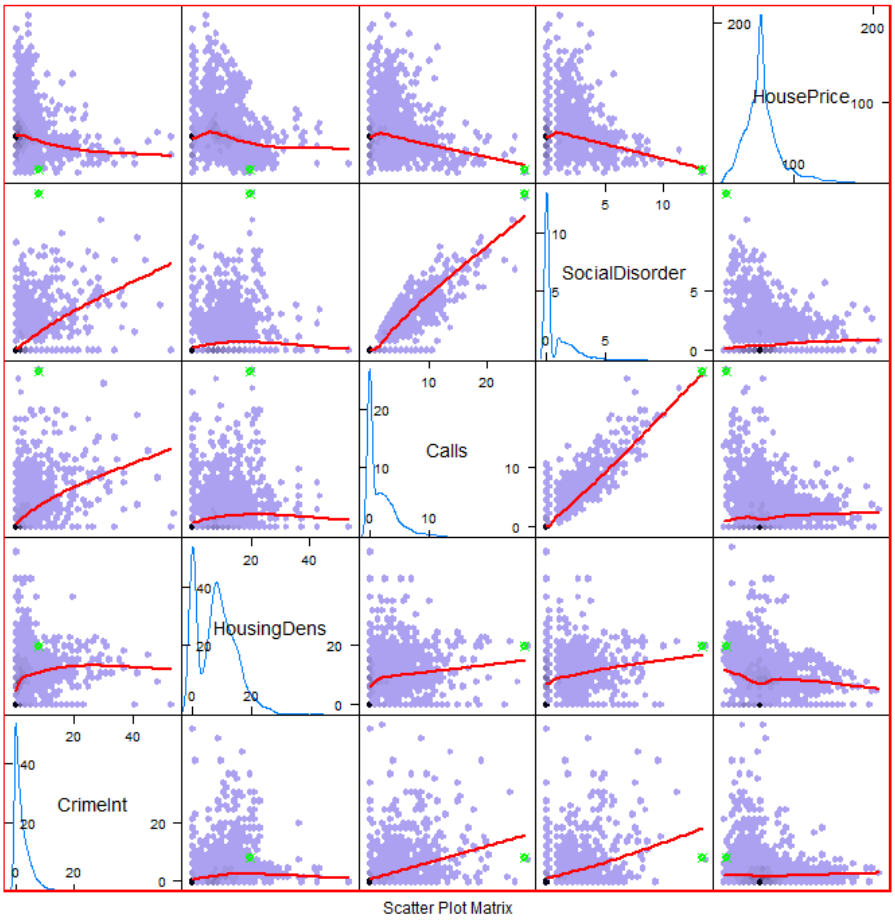

Figure 15 Scatterplot matrix with outlier identified.

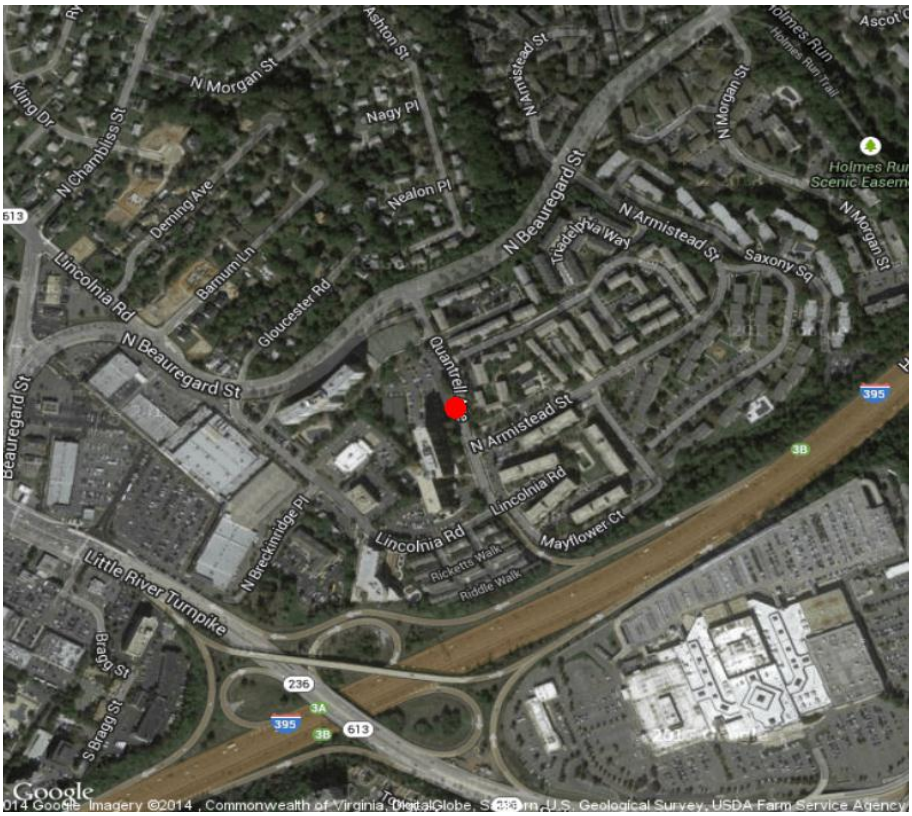

Figure 16: Close-up of outlier in full crime data set.

3.2.2 San Francisco Crime-Related Variables

I include San Francisco as a second location in my research to show my modeling is flexible to different geographies and variables. For San Francisco, elevation and speed limits are possible crime-related covariates. As you go down the large hill that makes up the center of the map of San Francisco into the downtown area in the Northeast, the crime tends to increase as shown previously in Figure 10. A video on this phenomenon is "The Joy of Stats" by Hans Rosling, where he drives down this hill and labels crime locations (Rosling, 2010). He also comments on how the San Francisco Police Department has made remarkable efforts to provide access to their crime data. Due to the way in which

cities develop, geographers document high correlations between economic wealth and elevation in many locations (New Orleans, Richmond, Atlanta, etc.), with the poorer parts of the city being located at lower elevations (Campanella, 2002). I obtained elevation at the block level from the City and County of San Francisco website data.sfgov.org. Speed limit data can also be obtained for this website for most major streets in San Francisco. This data is already in a polyline format and just needs to be divided up by intersection. As stated before, people tend to commit crimes along the paths that they normally take. Roads with high speed limits tend to be highways, with few crimes expected to be located on them. Figure 17 displays the speed limits in San Francisco, with a lot of the higher speed limits on the highways leading in and out of the city. The roads with lower speed limits tend to be smaller, more local roads, and may have higher crime on them. Many of the segments' speed limit information are not provided. However, in California the default speed limit for unlabeled roads is 25 mph; thus, 25 mph is imputed in place of all unknown segment values.

Figure 17: Speed Limits in San Francisco, CA.

I created similar scatterplot matrices with hexagon binning and smoothing line to that of the Alexandria variables with those I gathered for San Francisco. We see similar trends with crimes versus the Census crime-related variables (such as housing density) using hexagon binning, with perhaps an even stronger correlation. Higher elevation in Figure 19 is related to lower crime in the data set. It seems based on the smooth line that crime has a mostly stable trend across speed limits; however, the points themselves for speed limit do show some larger crime points for small speed limits. I will evaluate how

important each of the variables is in modeling the overall crime in San Francisco in Chapter 6.

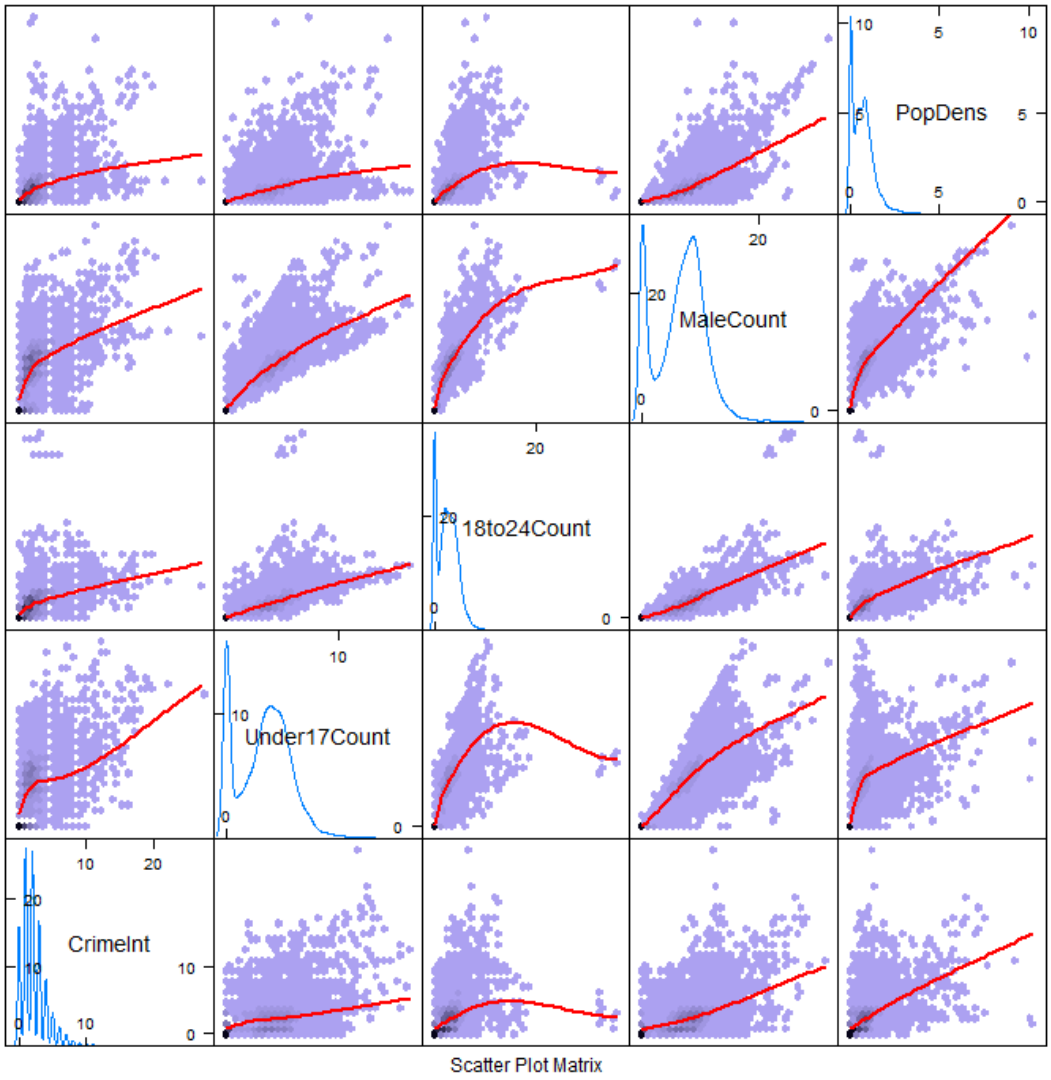

Figure 18: Scatterplot matrix with hexagon binning and loess smooth: San Francisco crime data set, first subset of variables.

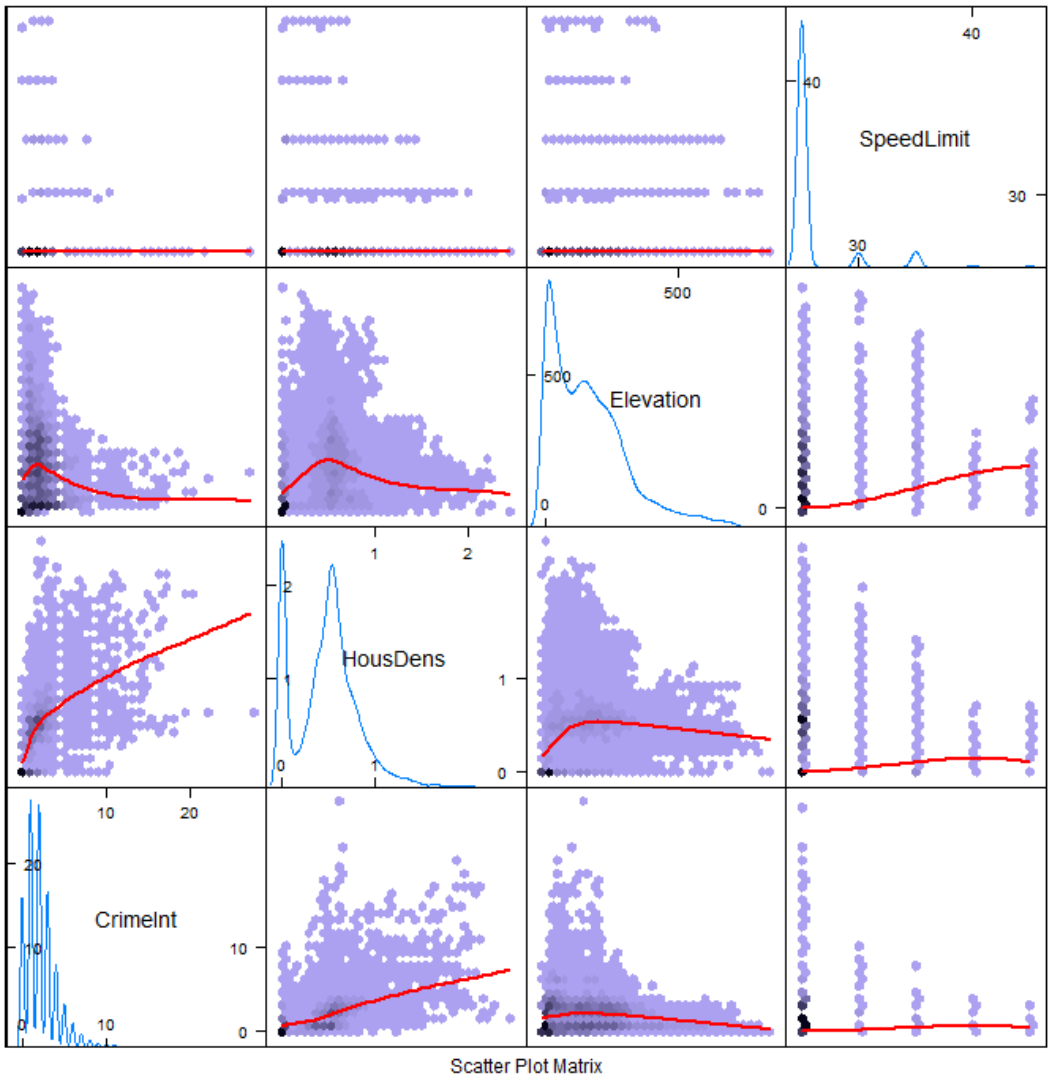

Figure 19: Scatterplot matrix with hexagon binning and loess smooth: San Francisco crime data set, second subset of variables.

CHAPTER 4. ASSIGNING AND SMOOTHING CRIME ON STREET SEGMENTS

4.1 Assigning area and point data to street segments

In order for street segments to serve as the fundamental units of crime analysis, methods or algorithms are needed that associate crimes and crime-related statistics with street segments. The area data consists of census blocks. It is straightforward to associate census block values to their bounding segments. Except for segments located along the boundaries of the map, each segment will get two values for a crime-related variable that represent the two sides of the street. In more general situations, it could be feasible that a road takes on three or more values. There are many different functions that could be applied to the two values to create one representative value for that segment. Here I use a simple average. Taking the difference between the two values could also be interesting for future analysis of differences between sides of streets.

I would like something more sophisticated to measure the crime points along the street segments. In the point data description below I refer to projecting crime events to street segments, since this is the dependent variable of primary interest. However, the same method is applicable to other variables assessed at "points". These could be either dependent variables or covariates in models.

When crime events are recorded, police or others produce reports that use street addresses (along with geospatial coordinates) as the crime location. In such situations

humans are basically making the association. Of course there are errors in reporting and subsequent processes; for example, more than one street could have the same name. I make use of the data provided knowing that there are some problems and provide one way of addressing the gap in associating crime points with street segments whose polyline vertices have known geospatial location.

Road segments are defined as the portions of road bounded by two intersections. A road segment is considered a polyline, which is a collection of line segments which are treated as one object. A block is a polygon, which is consisted of one or more polylines. I start with the case of crime located in a polygon whose edges are street segments. In order for street segments to serve as the fundamental analysis unit for crimes, methods with algorithms are need to associate point indexed data with street segments. Some algorithms are available in software such as ArcGIS for associating points with lines. A simple approach associates point data with the line to which it is closest. Geometrically, it is straightforward to assess the closest distance of a point to a line. The orthogonal projection of the point onto the line or an extension the line provides a basis for assessing distance. This is replaced by distance to the closest endpoint if the projected point is on an extension of the line. When the road segment is a polyline, the distance to the polyline can be the smallest of the distances to its constituent lines.

Associating a crime totally to the closest road segment seems a natural choice when the crime is much closer to the road segment than to other points of surrounding road segments. However, there are times when a crime can be almost equally close to two segments or more. This motivates assigning fractional crimes based on inverse

distance to surrounding road segments. The use of fractional crimes not may seem natural to those used to looking at discrete counts, but seems reasonable to address otherwise ambiguous situations. Also, accumulated fraction of crimes can be rounded to the nearest integer for presentation and discrete modeling purposes.

One approach to projecting crimes to segments is to allocate a fraction of the crime based on the relative distance of the nearest point of the each of the surrounding line segments. My approach allocates a fraction of the crime based on the relative distance to street segment midpoints. There were two reasons for this. First, the Census block boundaries are segments whose midpoints are more central to the blocks. Second, the segment midpoints offer a straightforward way assess the distance between connected segments. I use an inverse distance weighting so that crimes farther away from the road segment midpoint will receive less weight than those close to the road segment. Midpoint x will take a weighted value $u(x)$:

$$u(x) = \sum_{i=1}^{N} \frac{w_i(x)}{\sum_{j=1}^{N} w_j(x)} \tag{9}$$

where $w_i(x) = \left[\frac{1}{d(x,x_i)}\right]^{\beta}$ for $\beta \geq 0$ and $d(x, x_i)$ is the road distance from segment to crime point x_i, $i = 1,\ldots,N$. In my example I choose $\beta = 1$, which is simple inverse distance weighting. This transformation is monotone. As the values of β increase, weights given to points farther away from x decrease even more, while weights given to points close by increase more. If $\beta = 0$, each segment gets equivalent weighting.

In order to assign crimes from inside a polygon to a road segment, I compute these weights to assign each crime inside a polygon fractions of crimes for the segments

of that polygon. Figure 20 illustrates the algorithm one crime point inside a polygon with 5 segments. Each segment midpoint is assigned a "fraction" of the crime depending on how close that crime is to that midpoint. As stated previously, I can also reweight this with a different value of β, so that midpoints farther away get even smaller weights and the closer midpoints get even larger weights. The impact of a crime at a location inside a polygon of segments only has impact outside the polygon via segment distances from the polygon segments.

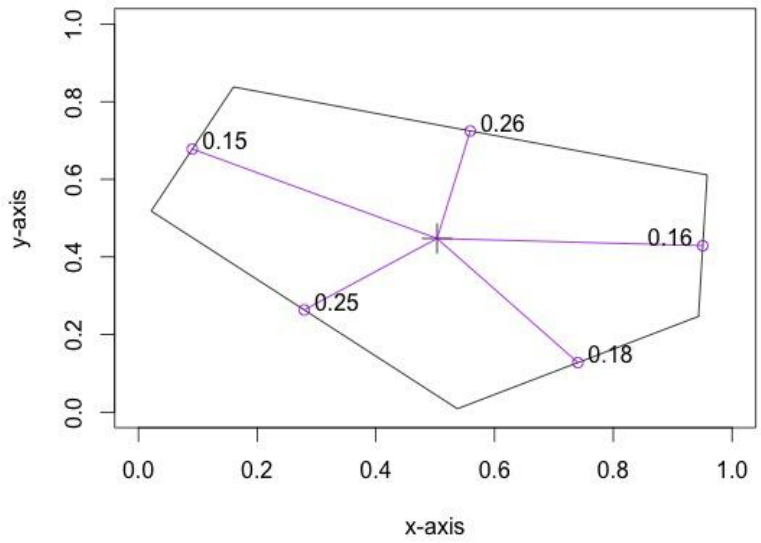

Figure 20: Weights of a Crime Assigned to Midpoints

The sum of all fractional counts at each midpoint is the crime value for that midpoint. Most segments have two sides of the street (two neighboring blocks/polygons) with fractional crime counts being accumulated from both sides. Note that there are some geometric concerns not considered when assigning crimes to road segments in this

way. An intuitive way to think about distance is the travel time for a human. Issues arise when there is a feature such as a building or river that is a barrier between a point and road segments. The closest segment (by travel time) may be different than the smallest distance to each segment.

More generally, additional methods of assigning points to segments and assessing the distance from segment to segment can be developed based on the topic and the kind of data available. Methods of assigning point data to road segments can be refined based on additional experience and scrutiny. For example, there may be times when barriers between points and lines exist that should be respected.

4.2 Smoothing crime counts over street segments

I smooth the crime counts over the road segments in order to better visualize the occurrence of crime hotspots. I would like the impact of a crime at a location inside a polygon of segments to have impact outside the polygon via segment-to-segment distance. I want to borrow strength from neighboring connected segment's crime counts to average out local spatial variation to reduce noise, but not so much that I oversmooth and increase bias. I measure the road distance from segment midpoint to segment midpoint in order to smooth the new fractional values of the crime counts over the road network space.

A segment is considered a "neighbor" to another is they both share at least one vertex. For each segment, I incorporate the values of crime counts from two levels of nearest segments to calculate new smoothed crime counts, as illustrated in Figure 21. I

consider all segments (blue) connected to the segment of interest (black), and all of the segments (red) connected to the blue segments. For the smoothing, I will use both the distance of the nearest segments and the angle at which they meet. The rationale behind using the distance is that I assume segments closer to the segment of interest will be the most similar in crime composition.

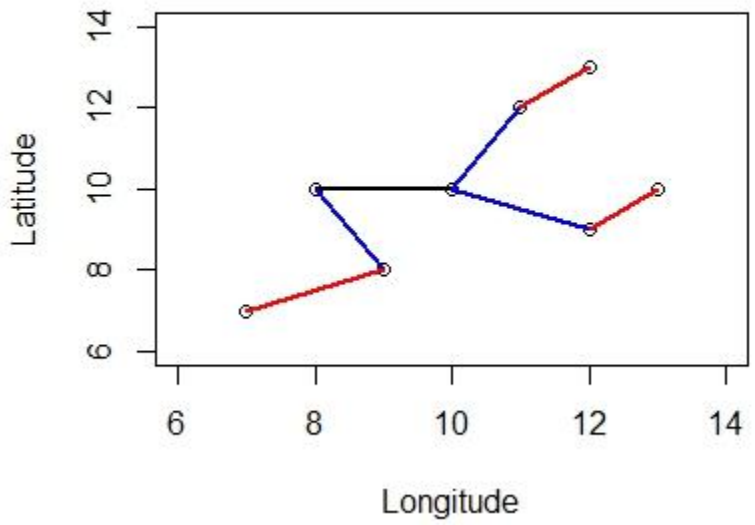

Figure 21: Graphic example of segments two levels apart. The black line is the original line, the blue lines are the nearest connected segments, and the red lines are the nearest connected segments to the blue lines.

For angles, I use a continuous scale that gives highest weight to those segments that create 180° angles at intersections. That is, the segments join in a straight line. I give the smallest weight to intersections that join with an 90° angle. High smoothing weights for segments making straight angles makes sense intuitively from a line-of-sight and traffic flow perspective—roads in a straight line to the next road would be expected

to be relatable. Similarly, intersections with small angles would be more related to each other since they are so close in area (small distance from one road to the next). There may be an increase in small walkways and alleyways between roads that meet at a sharp angle. The easier accessibility between roads that create sharp angles will make the roads more relatable in their crime counts.

The angle weight for all first-order neighbors will be produced using the angle between the original segment and those first-order neighbors. The angle weight for all second-order neighbors will be the angle between the second-order neighbor and its corresponding first-order neighbor. Figure 22 shows a graphical example of this. The black segment represents the original segment. When comparing with the red segment, the only angle under consideration is the one between the blue and red line.

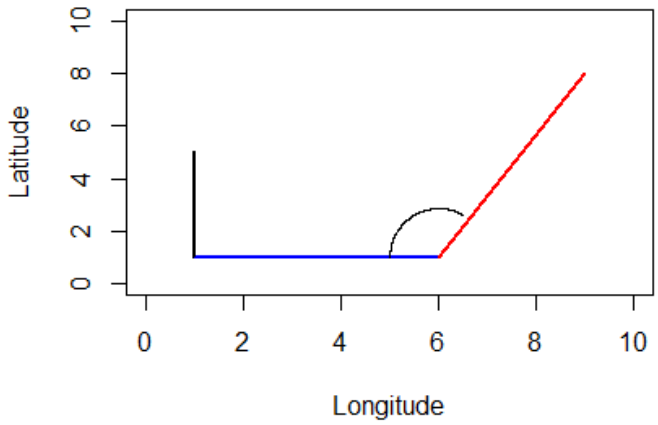

Figure 22: Depiction of angles: The black line is the original line, the blue line is the nearest connected segment, and the red line is the nearest connected segments to the blue line.

An alternative for calculating the angle weight of second-order neighbors could be to incorporate both the angle from the first neighbor and the angle of the second neighbor with a product of angle weights, but that was not done here. Looking at Figure 23, the angles considered when comparing the black segment to the red segment would be the one between the black and blue segment and between the blue and red segment. However, I did not use this angle weighting option in this dissertation.

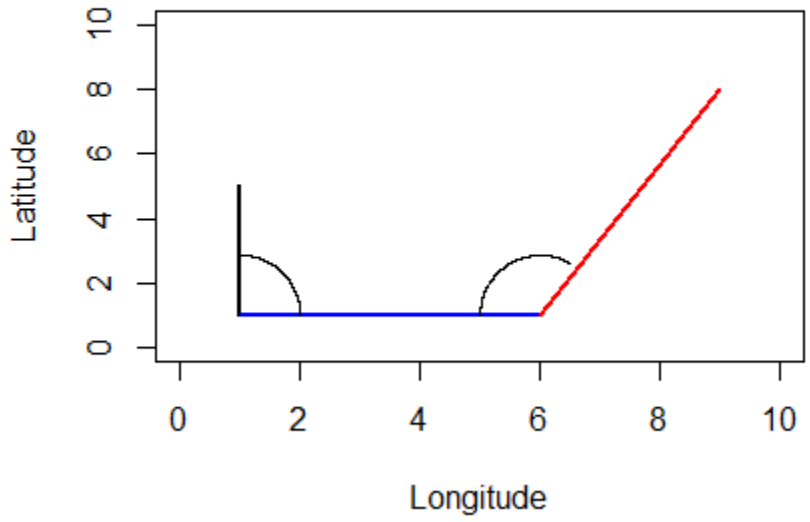

Figure 23: Alternative depiction of angles: The black line is the original line, the blue line is the nearest connected segment, and the red line is the nearest connected segments to the blue line.

I define the length from l_i from segment i (where i is in the set of neighboring segments $\{n(i)\}$) to the original segment to be the distance from the midpoint of segment

i to the midpoint of the original segment along the road network. I also define θ_i to be the angle at which each line segment i meets another at an intersection. Then the crime count of the original segment c_0 and the surrounding segments c_i for $i\epsilon\{n(i)\}$ are reweighted as follows:

$$c_{o_{new}} = ac_0 \tag{10}$$
$$c_{i_{new}} = (1-a)c_0 * \frac{w_i \frac{1}{l_i}}{\sum_{i\in\{n(i)\}} w_i \frac{1}{l_i}}$$

where c_i is the crime count of segment $i \in \{n(i)\}$ and $c_{o_{new}}$and $c_{i_{new}}$ are the new weighted values for the original segments and segment i, respectively. The angle weight w_i is defined as

$$w_i = b + \left[(1-b) * \frac{|90-\theta_i|}{90}\right]. \tag{11}$$

I define a to be how much weight is the fraction of the original segment count retained at that segment and $1-a$ to be the fraction of the original segment count attributed to its first and second segment neighbors. You could think about this as a type of broadcast smooth, with pieces of the original segment counts being given out to its surrounding segments. For example, for $a = 0.5$, half of the original crime count is incorporated into the smooth count for that segments and the other half is attributed to the first and second-level neighboring segments. The inverse distance weighting here gives smaller value to those segments whose distance is larger to the original segment. The value of b

determines how much weight is contributed by the angles of the surrounding segments. In my example I use $b = 0.5$, but choice of both a and b is flexible. Once $c_{o_{new}}$ and $c_{i_{new}}$ are calculated, they are summed up for each segment over the entire data set, creating the new smoothed counts.

This weight preserves the overall count of crime, while simply reassigning the values. When visualizing my smoothed counts, the length of segments may have a significant effect in analysis. Some roads are very long and have few intersections (such as highways), creating segments very long relative to most of the segments in the data set. These will then naturally higher counts to longer segments and lower counts to shorter segments. In order to standardize the values, I will divide the crime count for each segment by the length of that segment when mapping.

4.3 Smoothing Results

4.3.1 Alexandria Crime and Assaults

An algorithm was created in R which identifies the polygon (Census block) that each crime is in, along with the distance from each crime to each segment midpoint of that polygon. All of the inverse distances are calculated and each road segment midpoint is given the sum of these values. The data is imported in ArcMap of ArcGIS and each segment is given the value of its midpoint. Figures 24 and 25 give the resulting maps with a 5-point color scale with break points defined at quintiles for the full crime data set and the assault data set, respectively. I divide this fractional count value by the length so that roads don't get exorbitantly high values of crimes simply because they are long roads

with few intersections. The highest 20% of crime counts on the roads are shown in red. The big cluster of crime in Southeast Alexandria can still be seen.

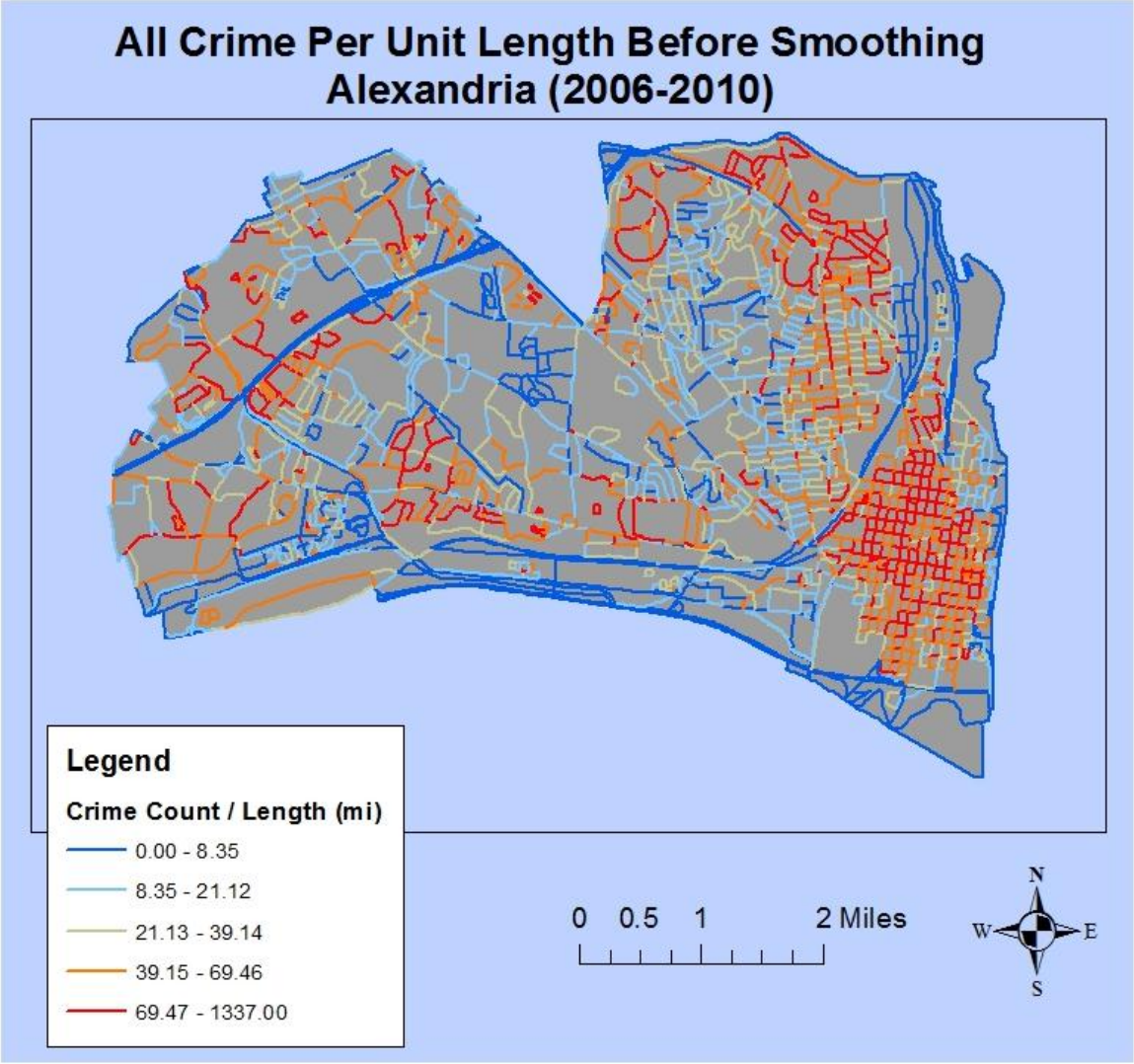

Figure 24: Crimes per unit length along road in Alexandria prior to smoothing

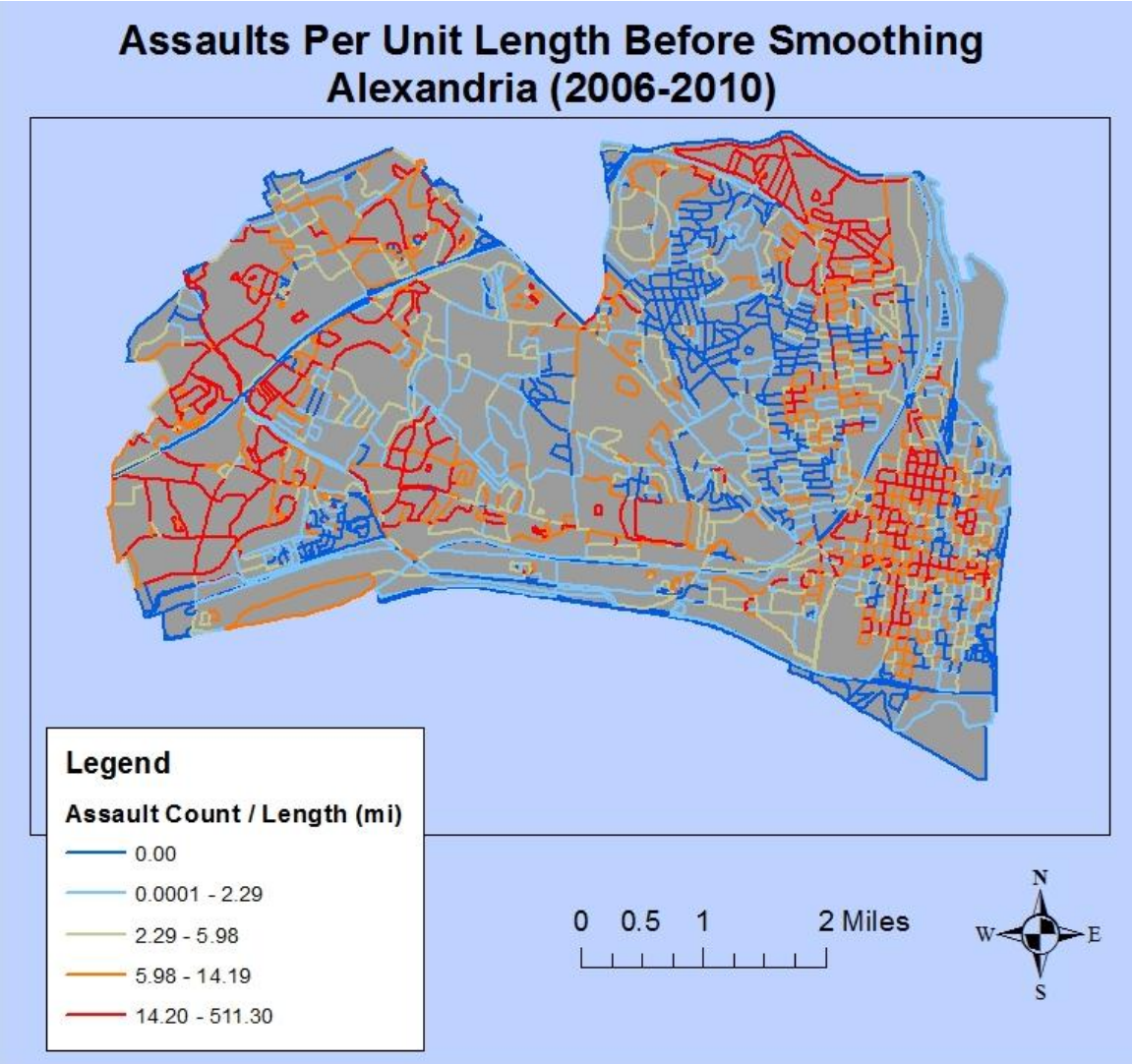

Figure 25: Assaults per unit length along road in Alexandria prior to smoothing

From Figure 25 with the Assault map you can see the 3 clusters defined previously (West, Northeast, and Southeast) even more clearly. In order to calculate the smoothed crime values, I take the fractional crime counts I have just compiled and smooth over the nearest roads, along with the nearest roads to those roads (two levels of nearest road segments). I do this by compiling a polyline shapefile in ArcGIS based on the census block polygon shapefile that includes the fractional crime count, the polylines

starting coordinates, ending coordinates, midpoint coordinates, and lengths. Note that converting the polygon to the polyline file creates duplicates of road segments except for those on the boundaries of the City of Alexandria. Using R, I sort the data by their coordinate values and join together the data that belongs to the same segment. Then, I created an algorithm that locates adjoining road segments by finding all segments that share starting coordinates and/or ending coordinates.

The distance between a segment and its adjoining segment is defined here as half of the length of the original segment added to half of the length of the adjoining segment. I then go another level, finding the adjoining segments to each segment already known to be connected to the original segment. The distance to the original segment here is the accumulation of half of the length of the original segment, the entire length of the adjoining segment, and half of the length of the segment connected to that. Angles between roads are also incorporated as described in earlier in this chapter. A list structure is compiled storing these lengths and angles, which I then use in the formula from Section 4.2 that gives some weight to the original segment and some weight to the neighboring list of segments.

Figures 26 and 27 show maps from ArcGIS using $a = 0.6$, meaning that 60% of the crime value comes from the original segment, while 40% come from its connected segments to second degree. The choice of a is up to the discretion of the user; it seems appropriate, however, that $a \geq 0.5$ so that at least half of the value of the segment is coming from its original value.

I keep the same break points on the color scale in the smoothed maps as those from Figures 24 and 25 for comparison. While the differences seem to be slight when comparing the two figures, you can see that some of the segments that were colored red in the previous figure in the North and the West are no longer considered as highly valued. Observe how the colors have a smoother transition from red to blue, as opposed to Figure 24 and 25 where the colors appear more scattered; this helps to identify more clearly the areas that may need more police patrolling. Many of the road segments in the Old Town Alexandria area remain red, as all of their nearest connected segments have equally high counts to the original segment and would be left relatively unchanged.

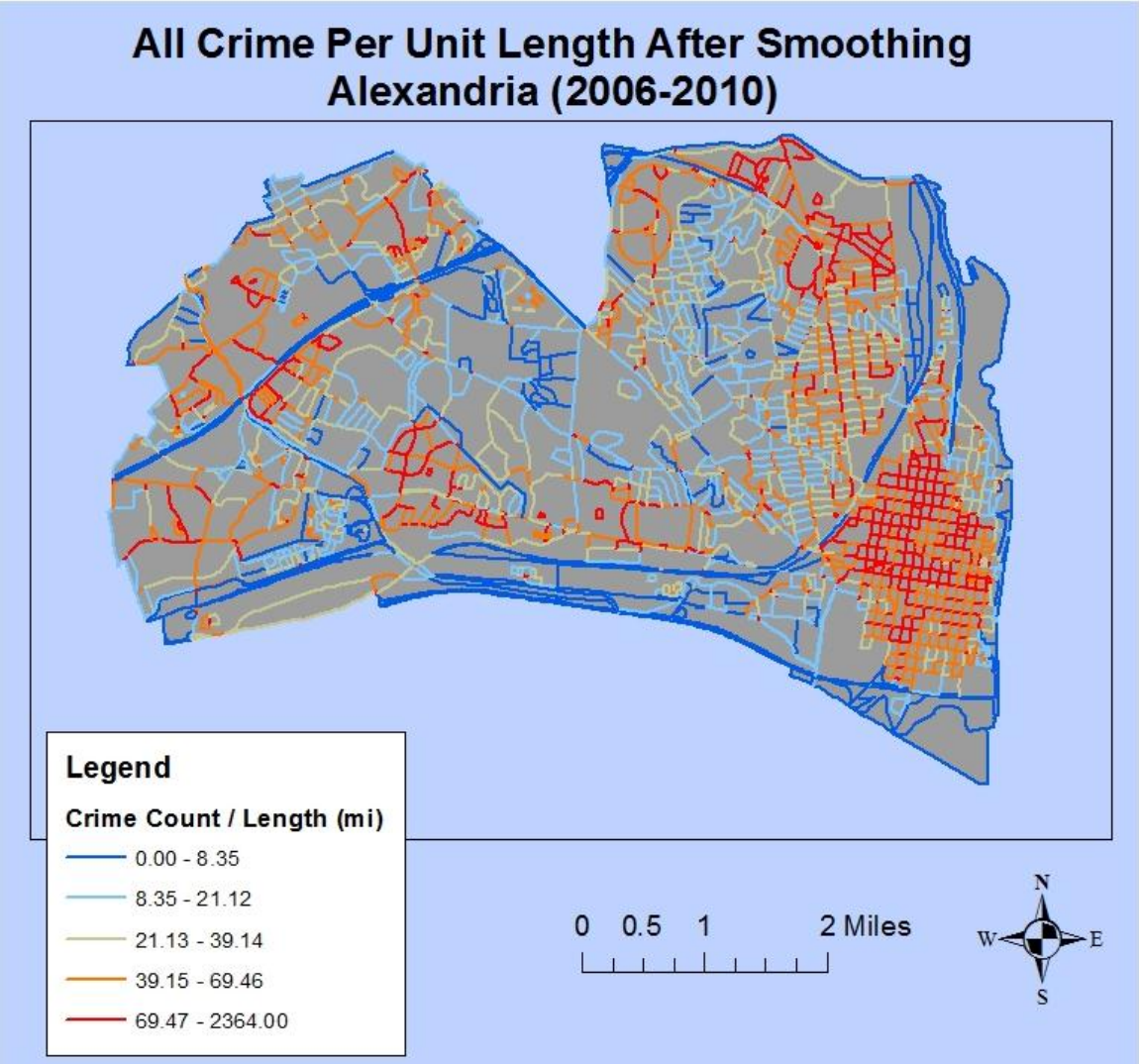

Figure 26: Crimes per unit length along road in Alexandria after smoothing, a=0.6.

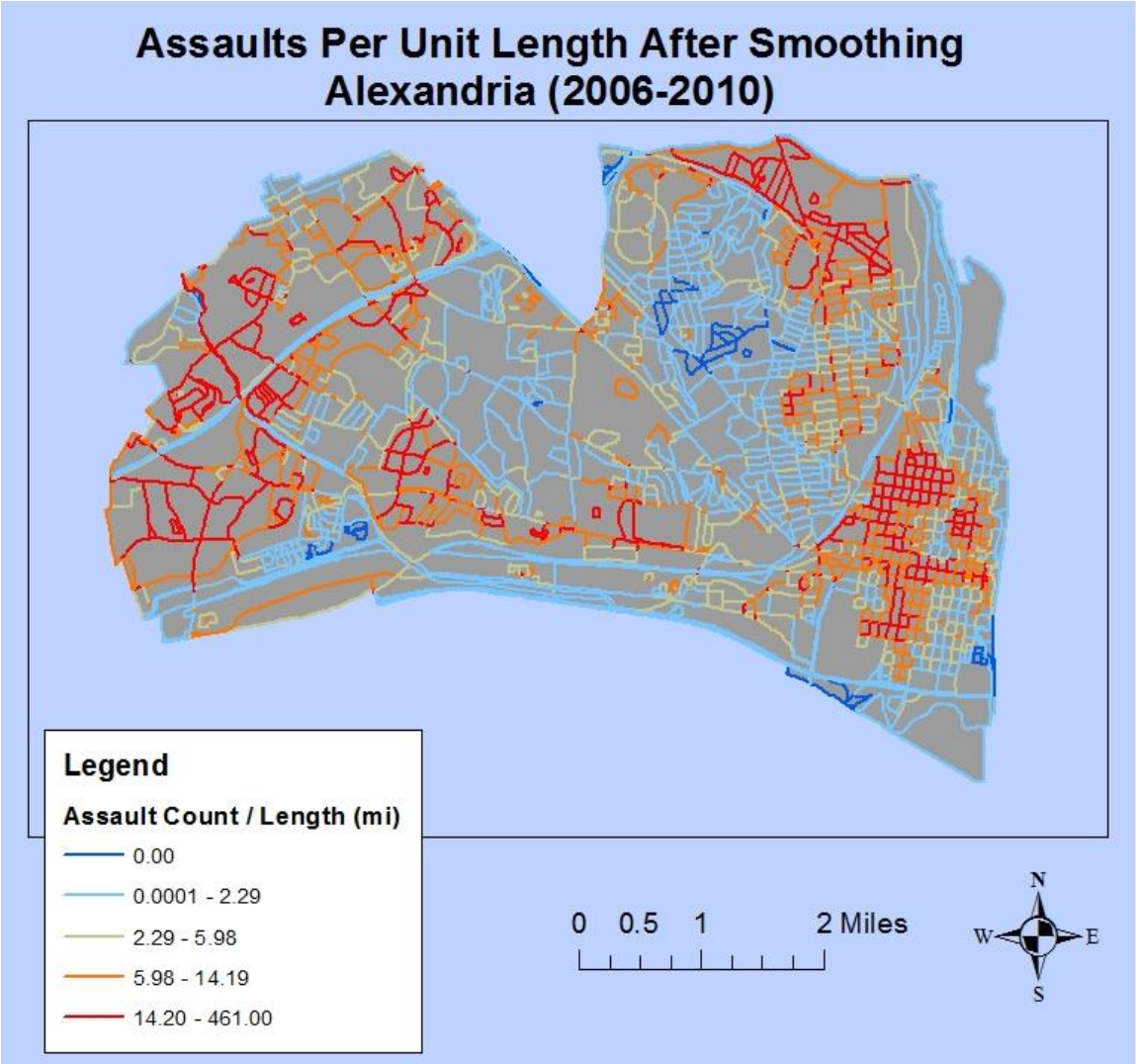

Figure 27: Assaults per unit length along road in Alexandria after smoothing, a=0.6

I compare Figure 27 with a different level of a in Figure 28. Here I use $a = 0.5$, which will use half of the value from the original segment and half of the value from neighboring segments. You can see very slight differences in these two figures.

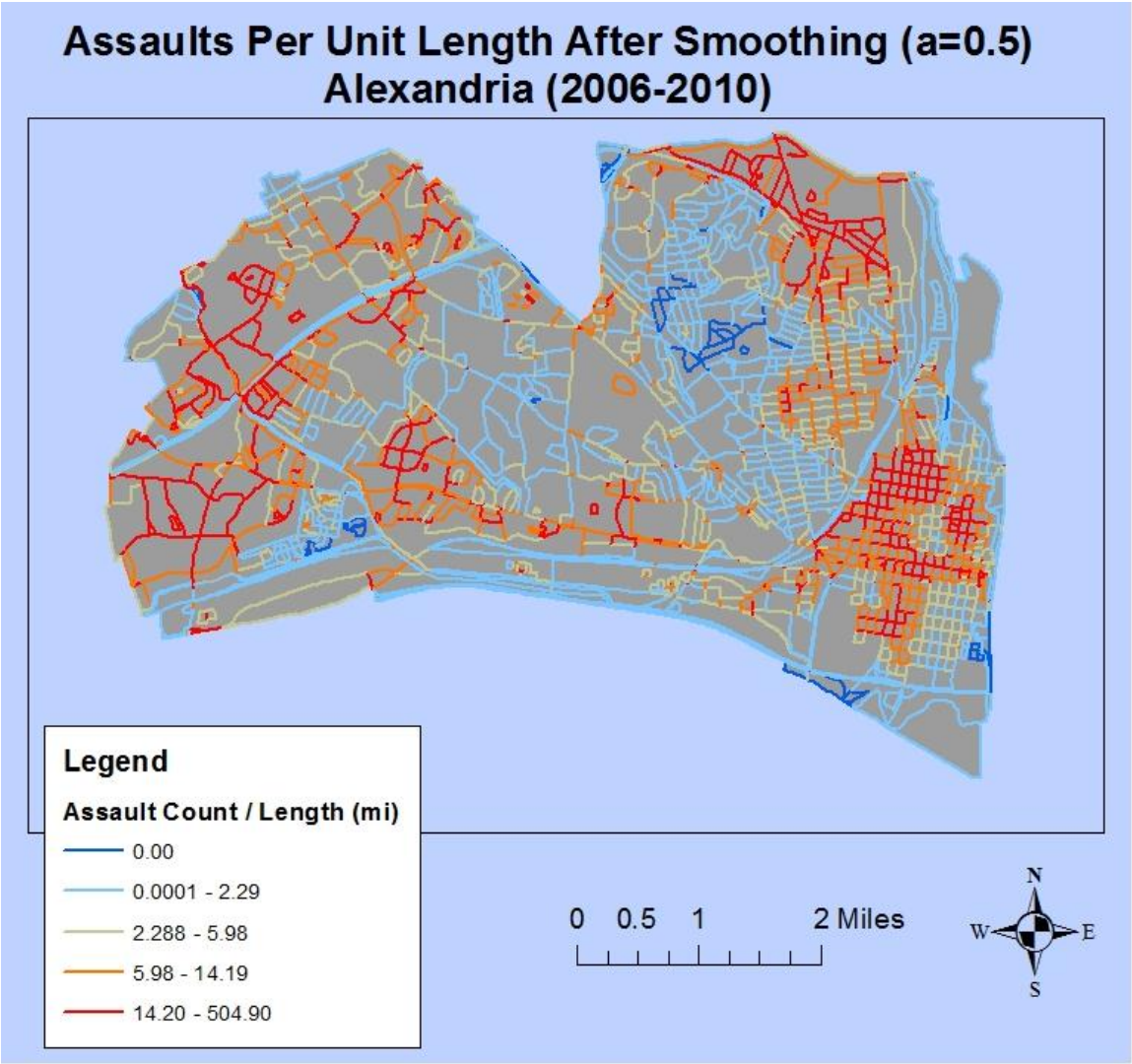

Figure 28: Assaults per unit length along road in Alexandria after smoothing, a=0.5.

4.3.2 San Francisco Crimes

I similarly look at crime maps for San Francisco, CA, once again with $a = 0.6$. You can see a large red patch in Northeast San Francisco, which is the downtown area, in Figure 27. Once again, red road segments indicate having the highest 20% of crime

counts compared with the other segments. Figure 28 shows the crime data after smoothing, with the values more smoothly transitioning from red to blue.

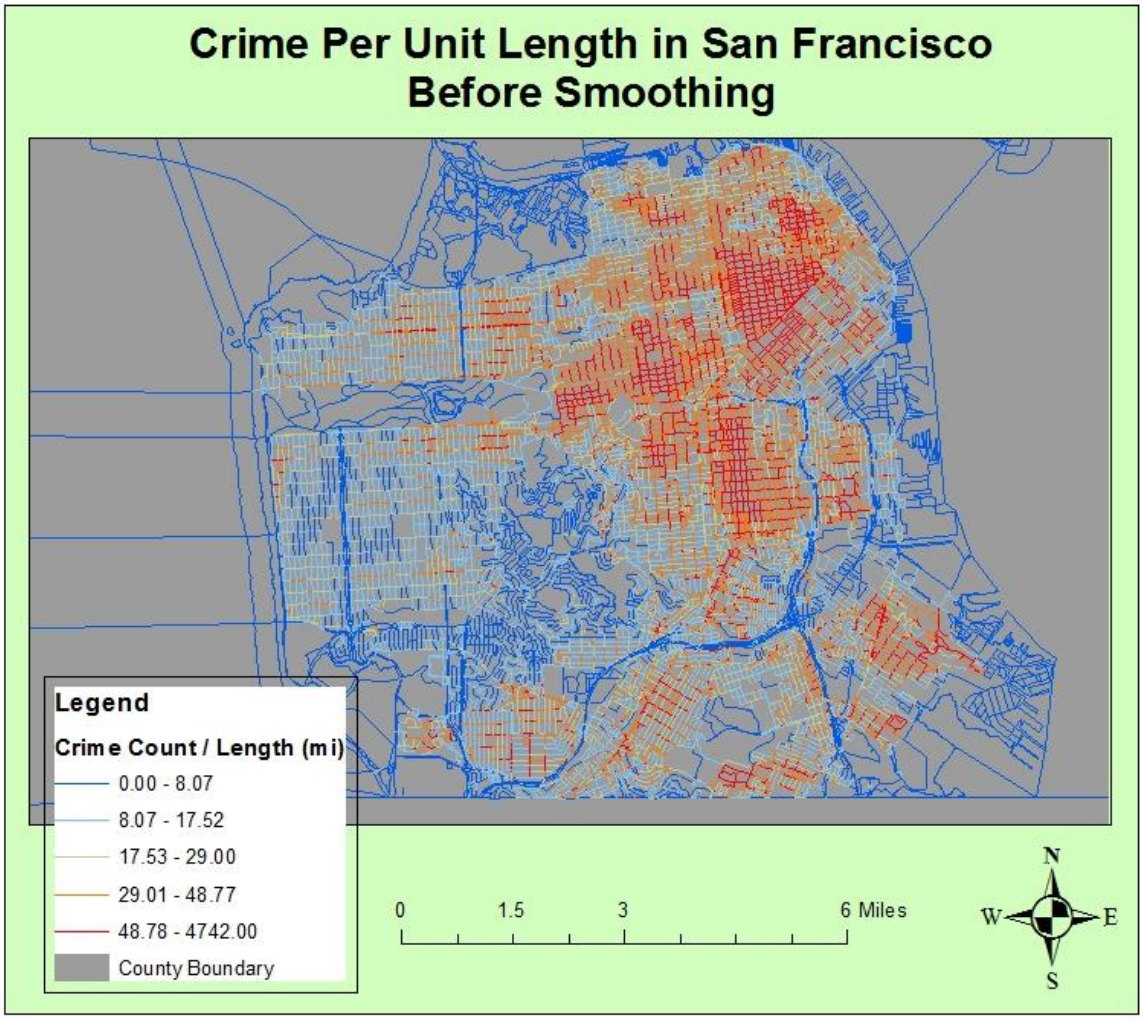

Figure 29: Crimes per unit length along road in San Francisco prior to smoothing.

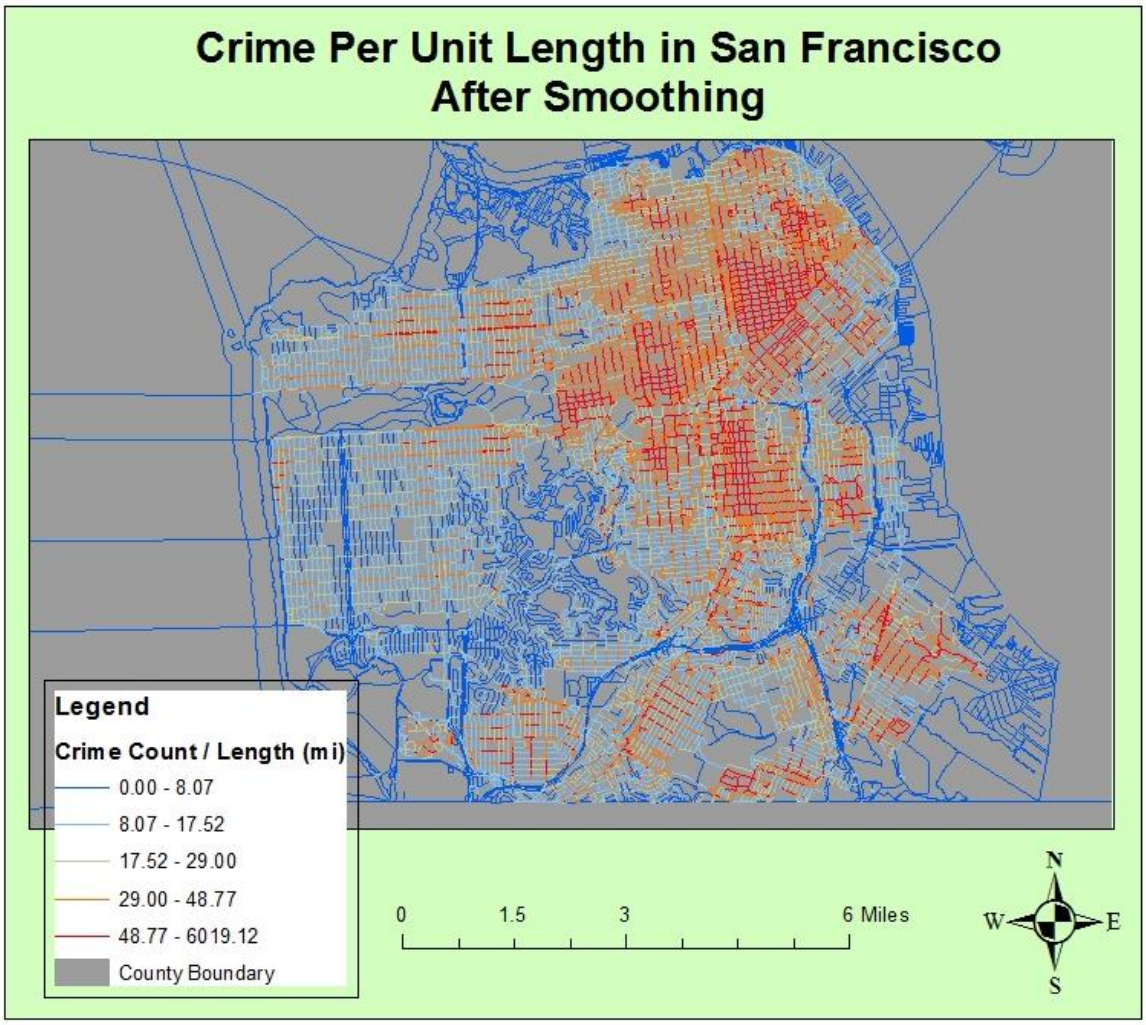

Figure 30: Crimes per unit length along road in San Francisco after smoothing, alpha=0.6.

CHAPTER 5. MODELING

5.1 Poisson and Negative Binomial (Zero-Inflated) Regression

5.1.1 Poisson Regression

The Poisson distribution is a discrete probability distribution that expresses the probability that a certain number of events occur in a given interval (Faraway, 2006). The Poisson distribution arises when the events being counted occur independently, the probability of two or more events being counted occurring simultaneously is zero, the events occur randomly in time or space, and the average count in an interval is proportional to the length of that interval. Formally, the probability mass function for random variable X with Poisson parameter $\lambda > 0$ is

$$\Pr(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}, k = 1,2, \ldots \quad (12)$$

A basic property of the Poisson distribution is that both its expected value (the average count in an interval) and variance are equal to a single parameter λ. In real-life examples, it may be the case that the variance is much larger than the mean. This is referred to as overdispersion (Cameron and Trivedi, 1998).

The Poisson regression model is the standard model for count data. This assumes that the response variable has a Poisson distribution and can be modeled by a linear combination of unknown covariates with regression coefficients $\boldsymbol{\beta}$. For a sample of n

independent Poisson random variables $y_1, y_2, \ldots, y_n$, a simple linear model with mean λ_i depending on explanatory variables $\boldsymbol{x_i}$ is

$$\lambda_i = \boldsymbol{x_i}'\boldsymbol{\beta}. \tag{13}$$

Since the left side of this equation, which is the expected count, must be nonnegative, it is typical to instead consider the log-linear model

$$\log(\lambda_i) = \boldsymbol{x_i}'\boldsymbol{\beta}. \tag{14}$$

Then for the i^{th} of n independent observations, the distribution of y_i given x_i is Poisson distributed with density

$$f(y_i|x_i) = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!}, y_i = 0,1,2, \ldots \tag{15}$$

where $\lambda_i = \exp(\boldsymbol{x_i}'\boldsymbol{\beta})$.

By the "law of rare events" (Levine et al., 2013), the total number of events will approximately follow a Poisson distribution if an event occurs in any of a large number of trials but the probability of occurrence in any given trial is small and assumed to be constant. Poisson regression is appropriate for the analysis of rare events such as crime incidents, motor vehicle crashes, and uncommon diseases. According to Haining (2003), the Poisson distribution is often used to model both crime counts and rare diseases.

5.1.2 Negative Binomial and Zero-Inflated Models

As stated in the previous section, overdispersion is a common problem when applying the Poisson distribution to real-life data. There are several different methods to account for overdispersion and to better model the data. Some software allows for

overdispersion in the fitting procedure. In some cases where the variance of the data is much larger than the mean, the more flexible negative binomial model, which has two parameters determining the mean and variance, may be a more appropriate choice (Faraway, 2006). Bayesian modelling can also address overdispersion by incorporation of random effects as well as fixed spatially structured components (Haining, 2003).

The standard form of the probability distribution for the negative binomial model is

$$\Pr(Y = k) = \binom{k + r - 1}{k} p^k (1 - p)^r \; for \; k = 0,1,2, \ldots \tag{16}$$

where p is the probability of success and r is the number of failures. The negative binomial distribution can alternatively be written as a mixture of Poisson and Gamma distributions:

$$\begin{aligned} f(k; r, p) &= \int_0^\infty f_{Poisson(\lambda)}(k) * f_{Gamma\left(r, \frac{1-p}{p}\right)}(\lambda) d\lambda \\ &= \int_0^\infty \frac{\lambda^k e^{-\lambda}}{k!} * \frac{\lambda^{r-1} e^{-\lambda(1-p)/p}}{\left(\frac{p}{1-p}\right)^r \Gamma(r)} d\lambda \\ &= \frac{\Gamma(r + k)}{k! \, \Gamma(r)} p^k (1 - p)^r. \end{aligned} \tag{17}$$

In my case, the y_i counts can be assumed to follow a Poisson distribution while the mean λ_i follows a Gamma distribution (Levine, 2013). Then the negative binomial distribution can then be defined as

$$y_i | \lambda_i \sim Poisson(\lambda_i) \tag{18}$$

with Poisson mean

$$\lambda_i = \exp(\boldsymbol{x_i}'\boldsymbol{\beta} + \boldsymbol{\varepsilon_i}) \quad (19)$$

where $\boldsymbol{\beta}$ is a vector of coefficients for the *k* parameters (plus an intercept) and the model error $\boldsymbol{\varepsilon_i}$ is independent with a gamma distribution with mean equal to 1 and variance equal to $1/\psi$ with $\psi > 0$ being called the inverse dispersion parameter (Levine, 2013). The negative binomial model can be useful not only when the variance is much larger than the mean, but also when the dependent variable is extremely skewed.

One of the possible causes of overdispersion in Poisson regression is the presence of excess zeros. Counts of zeros often appear in area-based count data when the populations are very small. As an extreme example, Loving County, Texas, with its 2010 population of 82, is very likely to have zero cases of death from leukemia in that year. There could be zero cases of deaths from any cause for a specific year. As indicated later, there is an excess of zero crimes in the Alexandria crime data both when treated as areal/block data and as road segments data. Also note that some blocks do not have any people living in them (population of zero). Out of the 1,294 blocks, a total of 347 of them have zero population. It might be helpful if such blocks were modeled differently. The negative binomial model in part helps to compensate for this problem.

More specific strategies to model crime data that have an excessive number of zero counts involve using the zero-inflated Poisson or zero-inflated Negative Binomial regression models, where the zeros are modeled separately. The theory behind zero-inflated models is that the zero count data is caused by a separate process and should be modeled independently of the other count data. Thus, the nonzero count data is modeled with the Poisson/Negative Binomial model while the zeros are modeled by a logit model.

Formally the zero-inflated Negative Binomial distribution is defined as follows (Cameron and Trivedi, 1998):

$$Pr(y_i = 0) = \varphi_i + (1 - \varphi_i)e^{-\mu_i}, \tag{20}$$
$$Pr(y_i = r) = (1 - \varphi_i)\binom{k + r - 1}{k} p^k (1 - p)^r \ , \ r = 1,2, \dots$$

where φ_i is the probability of extra zeros.

5.1.3 Alexandria Crime Modeling Over Area

For modeling over areas, the point data is summed up within each Census block, while the Census block data is kept in its original format. The R functions **glm()** and **zeroinfl()** in the **'pscl'** package can fit both simple Poisson and Negative Binomial regression models with or without the zero-inflated components. Here I apply the above models to the Alexandria crime data set and compare block-based models with road segment models with regards to modeling fits and residuals. I chose not to model San Francisco at this time because it has a much larger number of segments (over 16,000) and thus is much more computationally intensive. After exploring the data and making appropriate transformations, we put together the following models for analysis:

$$CrimeInt \sim Under17Count + Age18to24Count + (Age18to24Count * MaleCount) + MaleCount + PopDensity + HouseDensity + HousePrice + Calls + SocialDisorder$$

$$AssaultInt \sim Under17Count + Age18to24Count + (Age18to24Count * MaleCount) + MaleCount + PopDensity + HouseDensity + HousePrice + AssaultCalls + SocialDisorder$$

I will look at the full crime data set for the first four models. I will analyze both crime and assault models in the spatial models section. First, I calculate a simple Poisson regression model using the function **glm()** in **R**. The crimes are assigned to roads as explained in Chapter 4. These fractional crime counts are rounded to integers to be compatible with the discrete distribution model. Table 3 below shows the model results for the Alexandria, VA full crime data set. As stated before, I round crime to the nearest integer as these models will only take discrete values.

Table 3: Poisson Regression Model for Alexandria Crime Aggregated to Blocks.

Coefficients	Estimate	Standard Error	Z-Value	P-Value
Intercept	3.310	0.143	232.310	<0.001
Under17Count	0.060	0.002	25.889	<0.001
Age18to24Count	0.319	0.008	38.880	<0.001
MaleCount	-0.009	0.004	-23.155	<0.001
PopDens	-0.677	0.060	-11.264	<0.001
HousingDens	0.522	0.082	6.366	<0.001
HousePrice	-6.15E-4	2.25E-5	-27.296	<0.001
Calls	0.112	0.002	54.931	<0.001
SocialDisorder	-0.044	0.004	-9.499	<0.001
Age18to24*MaleCount	-3.23E-6	3.31E-7	-9.758	<0.001

Table 4: Deviance Residuals for Poisson Regression Model for Alexandria Crime Aggregated to Blocks.

Min	Q1	Median	Q3	Max
-36.267	-5.603	-3.024	0.180	103.416

The Poisson regression coefficient estimates, standard errors, and p-values are given in the output of the summary of the model. This model suggests that all of the variables have a significant effect on the crime count. However there are many indicators

suggesting that this model is not appropriate to use in this case. For Poisson models, the deviance residuals calculated in Table 4 should be approximately normally distributed if the model is specified correctly. Both the table with the median of -3.024 below zero and a large maximum and the boxplot (Figure 31) indicate the distribution is skewed to the right. The residual deviance in Table 5 can be used to measure the goodness of fit of the model. With the residual deviance of about 66319 and 1282 degrees of freedom, a chi-square test yields a small p-value (close to 0), indicating the data does not fit the Poisson model well.

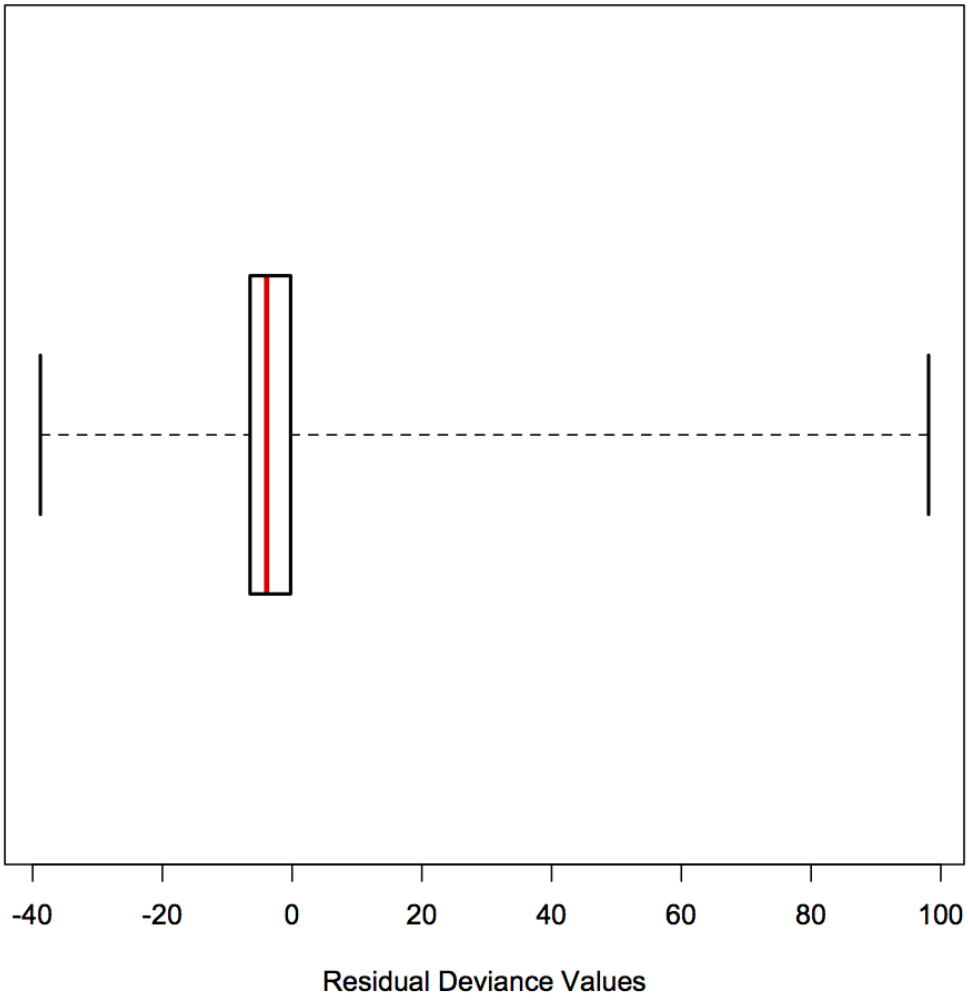

Figure 31: Residual Deviance Values from the Poisson Regression Model

Table 5: Residual Deviance of Poisson Model Aggregated to Blocks.

Residual Deviance	Degrees of Freedom	P-Value
66318.85	1282	<0.001

As explained previously, overdispersion can exist in Poisson crime count models like mine, with the variance being much larger than the mean. I use the function **dispersiontest()** in **R** to test the null hypothesis of no overdispersion and the p-value of

0.007745 appears in Table 6. One way to address this over-dispersion is using the Zero-Inflated Poisson model. A simple specification of this model is to assume all zero counts have the same probability of belonging to the zero component. I generate the following results using the **zeroinfl()** function, which once again gives that all of the variables have a statistically significant relationship to crime counts as shown in Table 7.

Table 6: Overdispersion Test of Poisson Model Aggregated to Blocks.

Z-Value	P-Value	Dispersion
1.736	0.041	166.658

Table 7: Zero-Inflated Poisson Regression Model for Alexandria Crime Aggregated to Blocks.

Coefficients	Estimate	Standard Error	Z-Value	P-Value
Intercept	3.897	0.015	264.50	<0.001
Under17Count	0.062	0.002	25.97	<0.001
Age18to24Count	0.254	0.005	54.76	<0.001
MaleCount	-0.052	0.002	-21.69	<0.001
PopDens	-0.912	0.063	-14.55	<0.001
HousingDens	0.402	0.084	4.82	<0.001
HousePrice	-0.001	0.000	-54.40	<0.001
Calls	0.108	0.002	54.40	<0.001
SocialDisorder	-0.056	0.004	-12.74	<0.001
Age18to24*MaleCount	-0.003	0.000	-25.50	<0.001

Similarly, I model this area data using the Negative Binomial and Zero-Inflated Negative Binomial Models. As shown in Table 8, the first simple negative binomial model yields very different results than the previous Poisson model. Many of the p-values have increased and some of the crime-related variables are no longer significant.

Specifically, this model gives the 18-24 year olds, population density, housing prices, and calls for service as having a significant effect on crime counts.

Table 8: Negative Binomial Regression Model for Alexandria Crime Aggregated to Blocks.

Coefficients	Estimate	Standard Error	Z-Value	P-Value
Intercept	2.752	0.100	27.636	<0.001
Under17Count	-0.020	0.041	-0.489	0.625
Age18to24Count	0.337	0.054	6.200	<0.001
MaleCount	0.015	0.032	0.459	0.646
PopDens	-2.032	0.705	-2.884	0.004
HousingDens	1.294	0.948	1.365	0.172
HousePrice	0.000	0.000	-2.186	0.029
Calls	0.169	0.027	6.250	<0.001
SocialDisorder	-0.007	0.056	-0.118	0.906
Age18to24*MaleCount	-0.003	0.002	-1.645	0.099

Table 9 shows the coefficients of the zero-inflated negative binomial model. Compared with the regular negative binomial model, while some of the p-values have gone down, the same four variables remain significant.

Table 9: Zero-Inflated Negative Binomial Regression Model for Alexandria Crime Aggregated to Blocks.

Coefficients	Estimate	Standard Error	Z-Value	P-Value
Intercept	3.173	0.090	35.090	<0.001
Under17Count	-0.032	0.035	-0.904	0.366
Age18to24Count	0.349	0.047	7.420	<0.001
MaleCount	-0.012	0.024	-0.531	0.595
PopDens	-2.108	0.571	-3.695	<0.001
HousingDens	1.303	0.752	1.734	0.083
HousePrice	0.000	0.000	-5.069	<0.001
Calls	0.157	0.022	7.060	<0.001
SocialDisorder	-0.023	0.045	-0.517	0.605
Age18to24*MaleCount	-0.001	0.002	-0.779	0.436

Model-fitting measures are used to calculate which model will give the best predictive values. In Table 10, I compare these past four models in how well they fit the data by comparing their mean squared error and AIC. The mean squared error is the average of the squared residuals from the model (the residuals are the difference between the observed values and the model's predicted values). AIC is short for Akaike's "An Information Criterion", which is calculated for one or several fitted model objects where a log-likelihood value can be obtained. It follows the formula $AIC = n * \ln(SSE) - 2\ln(n) + 2p$, where SSE is the residual sum of squares, p represents the number of parameters in the fitted model, and n is the sample size. The smaller the residual values and the smaller the AIC, the better the fit. The Poisson model gives wildly larger values, indicating that the Negative Binomial model, specifically the Zero-Inflated Negative Binomial Model, is the most appropriate.

Table 10: Model Comparisons for Alexandria Crime Aggregated to Blocks.

MODEL	MEAN SQUARED ERROR	AIC DF	AIC VALUES
POISSON	166.53	12	71368
ZERO-INFLATED POISSON	19.34	11	60437
NEGATIVE BINOMIAL	6.58	11	10628
ZERO-INFLATED NEGATIVE BINOMIAL	5.24	12	10595

5.1.4 Alexandria Crime Modeling Over Roads And Comparisons

I can run analogous models for a road segment data to see how this compares with area-based modeling. As I explain later, I assign fractional crime counts to the nearest road segments, aggregate them for each segment and round the result to integers. Table 11 shows a terse summary for the Poisson, zero-inflated Poisson, negative binomial, and zero-inflated negative binomial regression models.

Table 11: Model Comparisons for Alexandria Crime Assigned to Roads.

MODEL	MEAN SQUARED ERROR	AIC DF	AIC VALUES
POISSON	4.489	10	18334.83
ZERO-INFLATED POISSON	2.907	11	17773.81
NEGATIVE BINOMIAL	1.671	11	14422.27
ZERO-INFLATED NEGATIVE BINOMIAL	1.671	12	14424.27

Notice that the mean squared error calculations are much smaller than that of the area-based models. This is because the road-based models are calculating at a much finer scale, with 3,328 segments rather than 1,294 blocks. There is less room for large squared errors when modeling these smaller counts. When focusing on the AIC values, the Poisson regression road-based models give significantly better fits compared to the area-based model. However, this is not the case when comparing the negative binomial results, suggesting there are additional sources of unmodeled variation in the road segment models. The AIC values are much more stable across the four road-based models.

For the road segment data, the negative binomial models fit better than the Poisson models. Table 12 shows the coefficient results from the zero-inflated negative binomial model. Looking specifically at these results, the under 17 and 18-24 year olds, house prices, calls for service, social disorder, and the interaction term are significant. These results are slightly different than the area-based model.

Table 12: Zero-Inflated Negative Binomial Regression Model for Alexandria Crime on Roads Segments.

Coefficients	Estimate	Standard Error	Z-Value	P-Value
Intercept	0.482	0.057	8.407	<0.001
Under17Count	0.048	0.016	2.976	0.003
Age18to24Count	0.065	0.023	2.902	0.004
MaleCount	0.008	0.013	0.654	0.513
PopDens	0.012	0.011	1.212	0.225
HousingDens	0.001	0.014	0.056	0.955
HousePrice	-0.003	0.001	-3.991	<0.001
Calls	0.102	0.012	8.504	<0.001
SocialDisorder	0.087	0.024	3.602	0.000
Age18to24*MaleCount	-0.003	0.001	-5.222	<0.001

5.2 Spatial Models

As stated earlier, an assumption of the Poisson distribution is that the events occur randomly in time and space. The semivariogram based on residuals can be used as a guide to indicate if a model is needed with a more complicated variance structure based on spatial information (Schabenberger and Gotway, 2005). A semivariogram measures the average dissimilarity between data as a function of their separation in geographical space. At locations x and y over spatial field $Z(\cdot)$, the semivariogram $\gamma(x, y)$ is defined as

$$\gamma(x, y) = E(|Z(x) - Z(y)|^2)/2 \tag{21}$$

The semivariogram can suggest that there may be some spatial autocorrelation in the residuals.

A simple Moran's *I* test can also be used to assess spatial autocorrelation. Given random variable X and n spatial units (indexed by i and j), Moran's *I* is

$$I = \frac{n}{\sum_i \sum_j w_{ij}} \frac{\sum_i \sum_j w_{ij} (X_i - \bar{X})(X_j - \bar{X})}{\sum_i (X_i - \bar{X})^2} \tag{22}$$

where w_{ij} correspond to a matrix of spatial weights. Values close to -1 indicate spatial dispersion, while values close to +1 indicate spatial correlation in the data. A value of 0 indicates spatial randomness.

Statistical approaches to addressing the spatial dependence in data include adding a spatial component to the regression model. A spatial version of the Poisson model is known as the auto-Poisson model. Assume there are k observed quantities $z(s_i) = (z_1(s_i), \ldots, z_k(s_i))_i$ that are realizations of random variables at spatial location s_i which vary over *D*, a subset of the two-dimensional space (Cressie, 1993). The auto-Poisson model specifies a conditional probability given the value of random variable *Z* at neighboring locations $N(i)$ and incorporates an intensity parameter λ that is dependent on the space (Haining, 2003). Using the notation of Haining (2003) this spatial Poisson model is defined as:

$$P\{Z(i) = z(i) | \{Z(j)\}_{j \in N(i)}\} = \frac{{\lambda_i}^{z(i)} e^{-\lambda(i)}}{z(i)!} \tag{23}$$

where $\lambda_i = e^{\alpha(i) + \sum_{j \in N(i)} b(i,j) z(j)}$.

Here, $\{\alpha(i)\}$ is the set of site-specific effects while $\{b(i,j)\}$ represent between site interaction effects; it is assumed that $b(i,j) = b(j,i)$ and $b(i,j) \leq 0$ for all i and j. The non-spatial version will set $b(i,j) = 0$. Because one of the assumptions is $b(i,j) \leq 0$, the spatial model will only model negative spatial dependence; that is, it will only model competitive neighboring dependence (Haining, 2003).

Like others, I generalize this auto-Poisson model to include the regression variables pertinent to my study. Let μ_i denote the expected value of the response Y_i. Then for the Poisson distribution, $\mu_i = \lambda_i = e^{\alpha(i)}$. For the generalized linear model a link function is included, which is a function of μ_i set equal to a linear combination of parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$:

$$g(\mu(i)) = \beta_0 + \beta_1 X_1(i) + \beta_2 X_2(i) + \cdots + \beta_k X_k(i), i = 1, \ldots, n. \quad (24)$$

For the Poisson model, $g(\lambda_i) = \log(\lambda_i) = \alpha(i)$ (Haining, 2003). This model can similarly be constructed for the spatial version of the negative binomial model.

5.2.1 Conditional Autoregressive Models

While spatial autocorrelation can be modeled as given above, spatial structure may still remain in the residuals. More in-depth methods that are commonly used to represent spatial autocorrelation, specifically for non-overlapping areal units, are conditional autoregressive models. In conditional autoregressive models (CAR models), models are specified for the conditional probability distributions for each observation $Z(s(i))$ given the values of all of the other observations (Schabenberger and Gotway, 2005). It is assumed that $Z(s(i))$ depends only on another observation $Z(s(j))$ if the

location $s(j)$ is in some set of neighbors of $s(i)$, $N(i)$. This process is known as a Markov random field. Thus, with the conditional autoregressive approach, models are constructed for

$$f\Big(\, Z\big(s(i)\big) \mid Z\big(s(j)\big),\ s(j)\epsilon N(i)\Big). \tag{25}$$

The conditionally specified model described is an example of a hierarchical Bayes model. In a hierarchical model, the observed outcomes are conditional on a set of parameters which are also conditional on another set of parameters, or hyperparameters (Gelman et al., 1995). In these models there is not a spatial parameter $b(i,j)$ to estimate because the spatial dependence is not defined directly through the observations.

To estimate the parameter of interest for a given area in a study region, strength is borrowed other areas in the study region. An example of using this type of model using relative risk of disease is given in Haining (2003). Let $Z(i) = O(i)$ be the observed number of deaths of a certain disease observed in area i, with $i = 1, \dots, n$. $O(i)$ is independent and identically Poisson distributed with intensity parameter $\lambda(i) = E(i)r(i)$, where $E(i)$ is the expected number of deaths from the disease in area i and $r(i)$ is the relative risk of dying from the disease in area i. Then,

$$P\{O(i)r(i)\} = \frac{\big(E(i)r(i)\big)^{O(i)} e^{-E(i)r(i)}}{O(i)!}. \tag{26}$$

When different areas of a region have widely different population sizes, it would be advisable to choose an estimator for $r(i)$ so that there are more uniform levels of precision across the space. In Bayesian analysis, it is assumed that the $r(i)$ themselves

are random variables that follow their own probability distribution, known as the prior distribution. The flexibility of the model can also be used to choose a model where the $\{r(i)\}$ are spatially structured or unstructured.

Continuing to follow the notation of Haining (2003), let μ vary across areas $\{\mu(i)\}$ and define

$$\mu(i) = \beta_0 + \beta_1 X_1(i) + \beta_2 X_2(i) + \cdots + \beta_k X_k(i),\ i = 1, \ldots, n \tag{27}$$

where random effects are decomposed into spatially structured and unstructured components. The parameters $\beta_0, \beta_1, \beta_2, \ldots, \beta_k$ are drawn from another probability distribution, of which there are many different choices.

The negative binomial model that incorporates a spatial autocorrelation term is the MCMC Poisson-Gamma-CAR Model (Levine, 2013). The Poisson-Gamma-CAR model has three key properties: a Poisson mean, a Gamma dispersion parameter (similar to the negative binomial model), and an estimate of local spatial autocorrelation. This model is defined as $y_i|\lambda_i \sim Poisson(\lambda_i)$ where the mean λ_i of the model is defined by the distribution

$$\lambda_i = \exp(x_i'\beta + \varepsilon_i + \phi_i). \tag{28}$$

In this model, β once again is a vector of coefficients for *k* covariates plus an intercept term, and the model error ε_i is independent of these covariates. It can be shown that $\lambda_i \sim Gamma(\psi, \psi e^{-x_i^T \beta - \phi_i})$, where the prior distribution of ψ is a Gamma distribution with hyperparameters a_ψ and b_ψ (default values equal to 0.01 in CrimeStatIV to reflect a noninformative prior). $Exp(\varepsilon_i)$ is defined to follow a gamma distribution with mean 1 and variance $1/\psi$ for $\psi > 0$. The extra term here, ϕ_i, is the spatial random effect for

each observation; these spatial effects are assumed to be distributed as a multivariate normal model (Levine, 2013).

The CAR function, developed by Besag (1974), can be expressed as:

$$E(y_i|y_{j\neq i}) = g[\mu_i + \rho \sum_{j\neq i} w_{ij}(y_j - \mu_j)] \quad (29)$$

where g is a function related to the expected mean, μ_i is the expected value for observation i, w_{ij} is a spatial weight between the i^{th} observation and all other observations, and ρ is a spatial autocorrelation parameter that determine the size and nature of the effect of the neighborhood (Levine, 2013). The estimate of the spatial parameter ϕ_i from Equation 28 uses a function of this form. Then ϕ_i can be modeled (using notation from Levine 2013) assuming

$$p(\phi_i|\Phi_{-i}) \propto \exp(-\frac{w_{i+}}{2\sigma_\phi^2}[\phi_i - \rho \sum_{j\neq i} \frac{w_{ij}}{w_{i+}} \phi_j]^2) \quad (30)$$

where $p(\phi_i|\Phi_{-i})$ is the probability of a spatial effect with $w_{i+} = \sum_{i\neq j} w_{ij}$ (summed over all neighboring regions). Equation 30 is a conditional normal density with mean $\rho \sum_{j\neq i} \frac{w_{ij}}{w_{i+}} \phi_j$ and variance $\frac{\sigma_\phi^2}{w_{i+}}$. The parameter ρ determines the direction and magnitude of spatial effects, and w_{ij} is the spatial weight between neighboring regions i and j. From the variance term, $\sigma_\phi^2 = \frac{1}{\tau_\phi}$ (note that this term is the same for all observations). The parameter τ_ϕ is assumed to follow a Gamma distribution $\tau_\phi = \sigma_\phi^{-2} \sim Gamma(a_\phi, b_\phi)$. The hyperparameters a_ϕ and b_ϕ are each given a default value of 0.01 in CrimeStatIV to reflect a noninformative prior (Levine, 2013).

The spatial weight matrix **W** has off-diagonal entries w_{ij} and diagonal entries $w_{ii} = 0$. The matrix **D** is a diagonal matrix with elements w_{i+} on the diagonal (0 elsewhere). Then, as shown in Sun, Tsutakawa, and Speckman (1999), a ρ is chosen such that $\kappa_{min}^{-1} < \rho < \kappa_{max}^{-1}$ where κ_{min} and κ_{max} are the smallest and largest eigenvalues of $\mathbf{WD^{-1}}$, respectively, then Φ has a multivariate normal distribution with mean 0 and nonsingular covariance matrix $\sigma_\phi^2(\boldsymbol{D} - \rho\boldsymbol{W})^{-1}$.

Putting everything together, the parameters in the Poisson-Gamma-CAR model are $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_n)$, $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_J)$, ψ, $\boldsymbol{\Phi} = (\phi_1, \dots, \phi_n)$, τ_ϕ and ρ. Once this model is specified and initial parameter values are chosen, random samples can be drawn from the full conditional distributions of each parameter and the estimates for the coefficients are estimated based on the results of these samples. Inference in calculating statistics for this model is based on Markov Chain Monte Carlo (MCMC) simulation. An MCMC algorithm is an iterative tool that generates each sample based on the value of the previous sample. Ideally, the algorithm is run until convergence has been obtained. Initial values of an MCMC algorithm are usually chosen arbitrarily in software tools used to implement the model.

I address two different types of Poisson-Gamma-CAR models in the following section. The first of these is the standard area-based method explained previously, which compiles a matrix of 1s and 0s based on whether polygons are neighboring each other. I compare this with the CAR model that uses a dissimilarity metric and modify it to use it with road segments. The model proposed by Lee and Mitchell (2012) is based on the standard CAR prior with the restriction that ρ is fixed at 0.99 to ensure that there is strong

spatial smoothing globally, which can then also altered locally by using a function of the dissimilarity between areal units. They do this because large differences in the response are likely to occur when neighboring populations are very different. This dissimilarity is captured by a separate matrix, which could include social or physical factors such as the absolute difference in smoking rates or the proportion of the shared border that is blocked by some physical barrier that cannot be crossed (Lee and Mitchell 2012). I will modify this to use over road segments by creating two matrices: a matrix of 1s and 0s based on neighboring segments and another matrix with road distance between these neighboring segments. In this way, the structure of the road segments can be somewhat maintained and incorporated into the modeling.

CrimeStat IV (Levine, 2013) supports spatial modeling with Poisson and negative binomial models and CAR models, but major focus is on analysis of space over regions rather than over lines. The R package '**CARBayes'** can also be used to analyze data using various types of CAR models (Lee, 2013). Research that focus on modeling crime counts and traffic crashes using these CAR models similarly incorporate the spatial component over regions rather than on roads (Osgood 2000, Miaou et al. 2003 and Song et al. 2006). I adapt the models to roads.

5.2.2 CAR Modeling Results

The visually apparent clusters in the Alexandria maps strongly suggest spatial autocorrelation. I ran Moran's I test to see if the statistics support the visual impression.

The extremely small p-value indicates that there is indeed spatial autocorrelation, which means the data would be better suited to a model that incorporates a spatial component.

Table 13: Moran's I Test for Spatial Autocorrelation.

Statistic	Observed Rank	P-Value
0.1447	10001	<0.001

In **R** I look at **properCAR.re(),** which is a function that fits a Bayesian hierarchical model with spatially correlated random effects to the data, where the data likelihood can be binomial, Gaussian or Poisson. The random effects are modelled by the conditional autoregressive (CAR) model explained previously, with inference is based on Markov Chain Monte Carlo (MCMC) simulations. After the first 5,000 burn-in samples, I collected 25,000 samples for the modeling. This area (block data) model uses a modified prior proposed by Stern and Cressie (1999) and uses a weighting matrix of 1s and 0s based on nearest polygons. If the polygons (blocks) are touching they get a value of 1, otherwise 0. This matrix is then a 1,294 x 1,294 size matrix, with rows and columns representing each polygon. Table 14 shows area model results based on aggregating all crimes to blocks.

Table 14: Poisson-Gamma CAR model for Alexandria Crime Aggregated to Blocks.

Coefficients	Median	2.5%	97.5%

Intercept	3.4030	3.4029	3.4031
Under17Count	0.0376	0.0376	0.0377
Age18to24Count	-0.1453	-0.1454	-0.1451
MaleCount	-0.0822	-0.0822	-0.0821
PopDens	0.4149	0.4123	0.4151
HousingDens	3.8425	3.8398	3.8427
HousePrice	-0.0008	-0.0008	-0.0008
Calls	0.0891	0.0891	0.0891
SocialDisorder	-0.1839	-0.1839	-0.1838
Age18to24*MaleCount	-0.0217	-0.0217	-0.0217
Rho	0.59	0.38	0.74

As opposed to p-values, Bayesian modeling results include 95% credible intervals for the coefficient estimates. The table shows the 2.5% and 97.5% percentiles that provide these intervals. A way to interpret these values is to consider any variable with an interval that does not include zero as significant in the model. In this case, all of the variables in this model are significant and are related the response (crime count), as their 95% credible intervals do not include zero. The ρ calculated here is between 0.59 and 0.74, indicating the model has included some of the spatial autocorrelation.

Table 15 shows the area model results for the assault data. For the assault data set, population density and social disorder are no longer considered significant. The spatial autocorrelation has a wider interval in this case with ρ values between 0.40 and 0.92.

Table 15: Poisson-Gamma CAR model for Alexandria Assaults Aggregated to Blocks.

Coefficients	**Median**	**2.5%**	**97.5%**

Intercept	1.1174	0.7017	2.5259
Under17Count	0.0839	0.0709	0.1152
Age18to24Count	0.2763	0.1851	0.3766
MaleCount	-0.0447	-0.1083	-0.0174
PopDens	-0.2284	-0.9247	0.0277
HousingDens	-1.0694	-3.4912	-0.6983
HousePrice	-0.0010	-0.0011	-0.0009
AssaultCalls	0.1932	0.1743	0.2314
SocialDisorder	-0.0072	-0.0244	0.0168
Age18to24*MaleCount	-0.0116	-0.0177	-0.0074
Rho	0.78	0.40	0.92

For the road segment data, I use a modified CAR model with a dissimilarity metric based on the length of all of the nearest segments. I create a matrix for all pairs of nearest segments that captures all of the distances between those segments and use this as the dissimilarity metric. I also include a matrix similar to the area-based model, with 1s and 0s based on nearest segments (1s if roads share an intersection, 0s otherwise). Both of these matrices are then 3,328 x 3,328 size matrices, with rows and columns representing each segment. I model the data using the R package **dissimilarityCAR.re()**. Tables 16 and 17 show results for the full crime-based model and the assault-based model, respectively. With the size of the matrices here, the run time of the code increases, with full-crime model taking approximately 40 minutes to run.

Table 16: Poisson-Gamma CAR dissimilarity model for Alexandria Crime Assigned to Roads.

Coefficients	**Median**	**2.5%**	**97.5%**

Intercept	0.0814	-0.0588	0.1934
Under17Count	0.0772	0.0526	0.1094
Age18to24Count	0.0518	-0.0003	0.121
MaleCount	0.0280	-0.0001	0.0568
PopDens	0.0030	-0.0163	0.0204
HousingDens	0.0073	-0.0161	0.0343
HousePrice	-0.0028	-0.0045	-0.0013
Calls	0.0822	0.0540	0.0992
SocialDisorder	0.1020	0.0644	0.1616
Age18to24*MaleCount	-.0054	-0.0067	-0.0039

Table 17: Poisson-Gamma CAR dissimilarity model for Alexandria Assaults Assigned to Roads.

Coefficients	**Median**	**2.5%**	**97.5%**
Intercept	-1.1929	-1.4702	-0.8196
Under17Count	0.0645	0.0120	0.1148
Age18to24Count	0.2586	0.1693	0.3461
MaleCount	-0.0623	-0.0120	-0.0212
PopDens	0.0057	-0.0308	0.0517
HousingDens	0.0043	-0.0560	0.0520
HousePrice	-0.0145	-0.0199	-0.0100
AssaultCalls	0.2234	0.1415	0.3144
SocialDisorder	0.1565	0.0855	0.2074
Age18to24*MaleCount	-0.0033	-0.0055	-0.0007

In the full-crime data set results in Table 16, the significant variables include the under 17 year olds, housing price, calls, and social disorder. While the 18 to 24 year olds were not significant, the interaction between this age group and males does have a significant effect in the model. In the assault model results (Table 17), population and housing densities are the only two variables not considered to have a significant effect on assault crimes.

I can compare the road-based models with the previous area-based CAR models. The model fitting calculations I used here are the mean squared error and the DIC. The

deviance information criterion (DIC) is a hierarchical modeling version of the AIC and is specifically useful in Bayesian model selection problems such as mine. For the area-based model, the R packages had some limitations so did not produce the model-fitting values. To address the problem, I ran the same model in CrimeStat IV.

For the area-based full crime model, The DIC measure is 47369, which is better than the AIC for the Poisson regression but worse than the negative binomial values. I also calculate a mean squared error of 1.01E+7, which is significantly larger than the mean squared error of the previous models. The reason for this is a number of large residual outliers. Below is a table of summary statistics which show a minimum residual value of -73050, which obviously has a major effect on the mean squared error. However, even locating and removing the outliers in R while yield similarly large values. For the area-based assault model, the DIC value from CrimeStat IV in this case is 7019, which is smaller than the AIC of the zero-inflated negative binomial model. The mean squared error in this case is much more manageable than the full crime data set at approximately 29258. Since we are dealing with a smaller number of counts overall in the assault data set, the errors will be smaller relative to that.

Table 18: Summary Statistics for Residual in Area-Based Full Crime Model

Minimum	Q1	Median	Mean	Q3	Maximum
-73050.00	-1.92	1.83	-186.30	15.87	1605.00

The following table gives a comparison of the area-based and road-based model-fitting results. The road-based CAR model has smaller DIC and mean squared error values than the area-based model. The segment-based model has a much larger number of segments and is looking at smaller counts of crime over a much finer level, leading to much smaller errors for better prediction along each segment.

Table 19: Comparing Mean Squared Error and DIC measures

MODEL	MEAN SQUARED ERROR	DIC
AREA-BASED ALL CRIME	10125263.74	47369
ROAD-BASED ALL CRIME	1.10	12957
AREA-BASED ASSAULTS	29258.22	7019
ROAD-BASED ASSAULTS	0.33	5834

Multicollinearity of the explanatory variables complicates interpretation of regression parameters and inflates standard errors. This is a serious issue with the data and could produce strange effects. Multicollinearity exists when two or more explanatory variables in a multiple regression model are highly linearly related. The correlations between these independent variables are strong. This could lead to unreliable and unstable estimates of regression coefficients. In these cases, the variable with the stronger correlation may have the correct sign while the weaker one will sometimes get flipped around. Another problem could be that the two variables cancel each other out. Sometimes the regression model will break down because the covariance

matrix cannot be inverted. High crime communities usually have many factors that might cause crime, and tend to be found in the same places at the same time. Thus a number of possible causal factors could be highly correlated with each other and multicollinearity becomes an issue (Vold, Bernard and Snipes, 2002).

A popular diagnostic tool for measuring multicollinearity is the variance inflation factor (VIF). This is calculated for each predictor by doing a linear regression of that predictor on all the other predictors, and then obtaining the R^2 from that regression. The VIF is then $\frac{1}{(1-R^2)}$, with strong multicollinearity being represented by VIF values far away from 1. Just looking at the simple negative binomial case, we get the following variance inflation factors for each variable:

Table 20: Variance Inflation Factors for Variables in the Negative Binomial Model for the Full Crime Data Set

Variable	Variance Inflation Factor
Under 17 Count	11.91
Age 18 to 24 Count	16.21
Male Count	20.58
Population Density	26.53
Housing Density	26.57
House Price	1.10
Calls	4.65
Social Disorder	4.55

Age 18 to 24 * Male Count	10.09

The resulting variance inflation factors show that many of the variables are highly correlated with each other. This doesn't necessarily affect overall model prediction, but it may lead to unreliable interpretations of whether individual variables are truly significant or not. As will be shown in the following section, in spite of the multicollinearity, the segment-based model will give a very decent prediction of crime compared with the area-based version. Further investigation into important variables and variable selection is given in Chapter 6.

5.3 Alexandria Crime Prediction over Area and Roads

I would like to see how well these models predict crime from one year to the next to see if the road-based methods predict crime better than the area-based methods. Using quantitative methods to predict where crime is located is known as "predictive policing" (Ratcliffe, 2004). In predictive policing, analytical techniques identify likely targets for police intervention, preventing crime and helping solve past crimes by making statistical predictions. Challenges of predictive policing include obtaining relevant, reliable data sets and having computing resources for computationally intensive algorithms.

I use my CAR-based models to analyze the data from 2006 to 2008 in order to see how well each method predicts the crime from 2009. I average the crime count values over the 3 years (2006-2008) and model these values with the two methods. An

alternative would be to smooth the crime counts from these years for comparison. Note that the covariates used in this research are the same for all three years. The modeling would be more complex if year-specific covariates were involved. I take the average difference between the fitted data from the modeling results and the true observed data from 2009 in Table 21. Based on the results in Table 21, the prediction values for the road-based model are very close to the true values for 2009. It also looks like the prediction in the road-based model has much less variability when predicting future crime than the area-based model. For example, the first and third quartiles for the area-based full crime data set are -3.30 and -0.90, while for the road-based it is -2.69 and -0.26.

Table 21: Summary of difference between Predicted and Observed 2009 values for Four Models.

Model	Min	Q1	Median	Mean	Q3	Max
Area-Based All Crime	-309.40	-3.30	0.016	30.32	0.90	14530.00
Road-Based All Crime	-38.04	-2.69	-1.20	-2.08	-0.26	0.96
Area-Based Assaults	-19.53	0.04	0.05	2.67	0.10	1098.00
Road-Based Assaults	-10.17	-0.12	0.01	-0.17	0.02	3.99

The following four figures represent the resulting maps using lines and areas based on predicted values from the models along with the observed crime points. It can be seen in these figures that the road-based model gives a more accurate prediction of future crime than the area-based model at a more local level. Figure 36 gives the difference between the predicted and the observed crime counts along the roads. You can

see there are more segments that underestimate the crime rather than overestimate. However, most of the segments seem to very closely predict the crime in 2009.

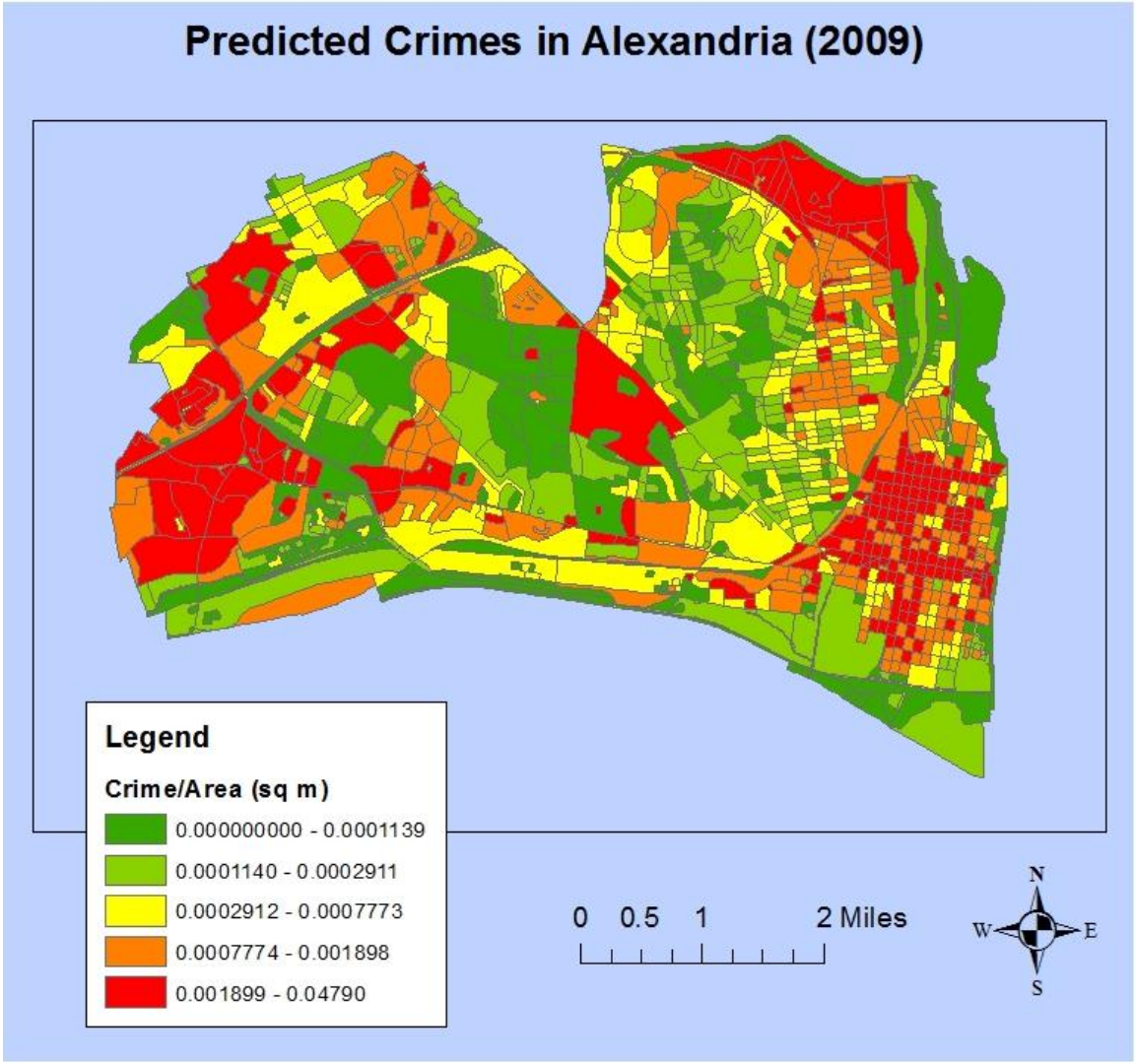

Figure 32: Predicted crime values over areas for 2009 in Alexandria, VA.

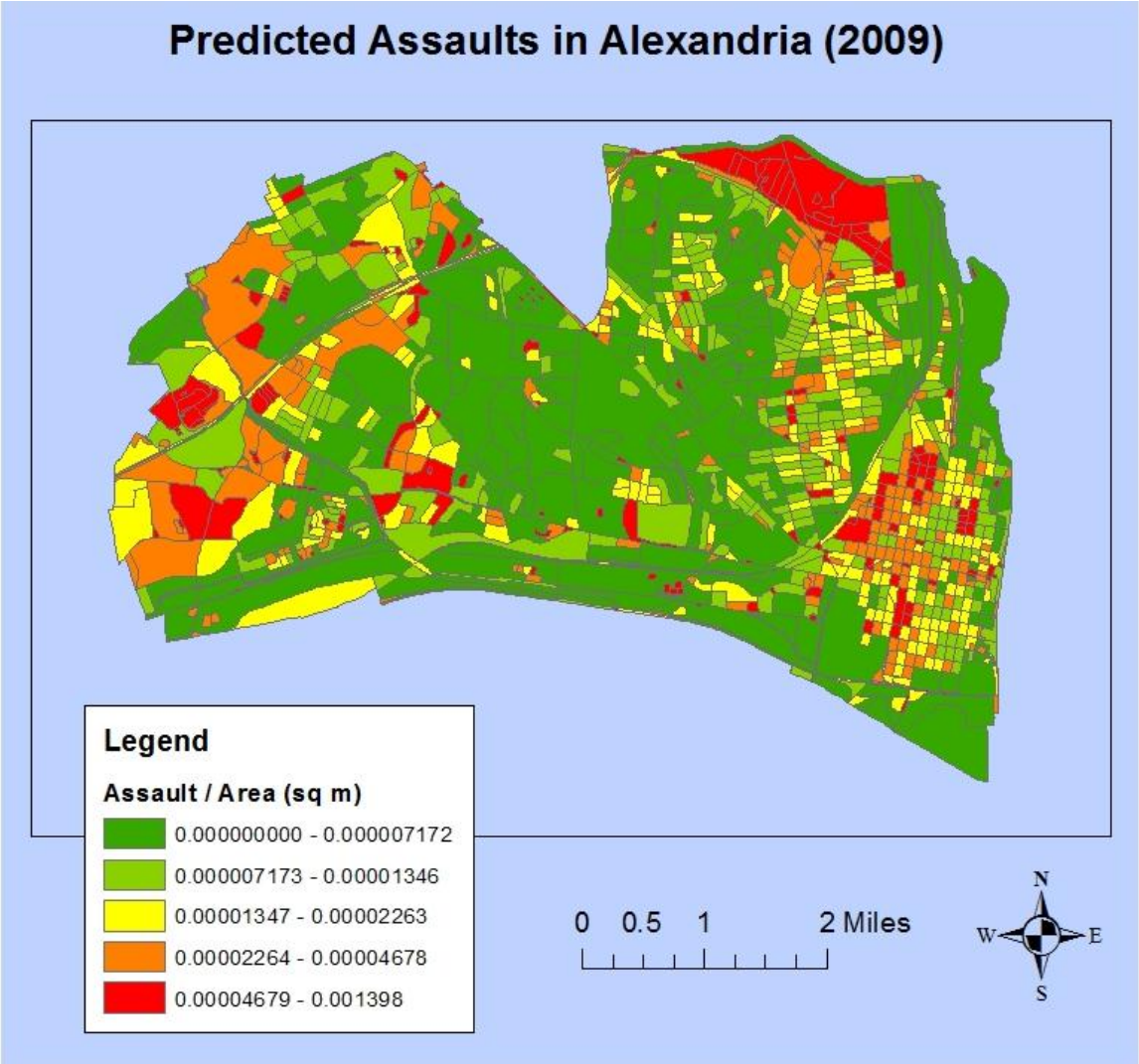

Figure 33: Predicted assault values over areas for 2009 in Alexandria, VA

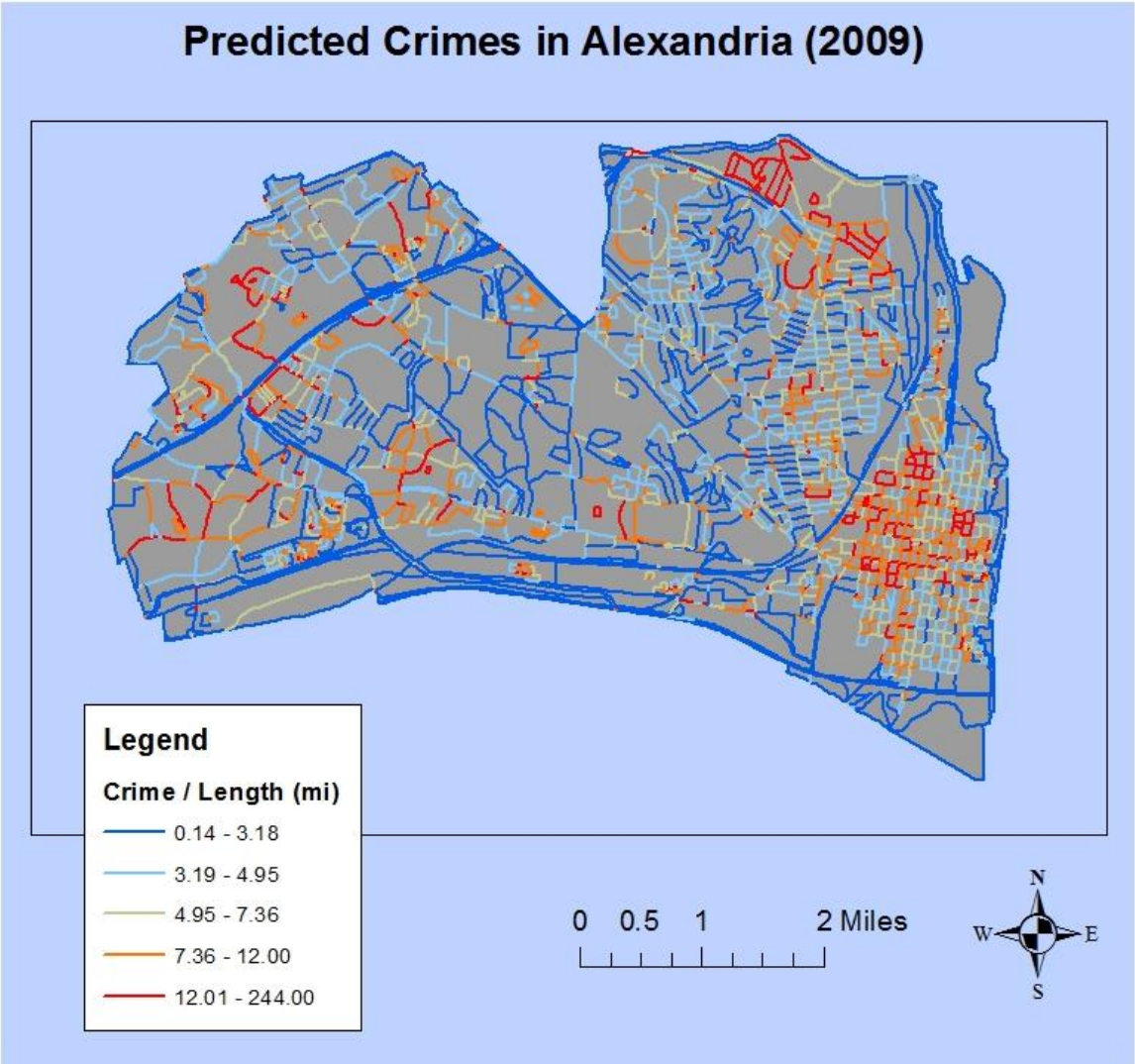

Figure 34: Predicted crime values over polylines for 2009 in Alexandria, VA

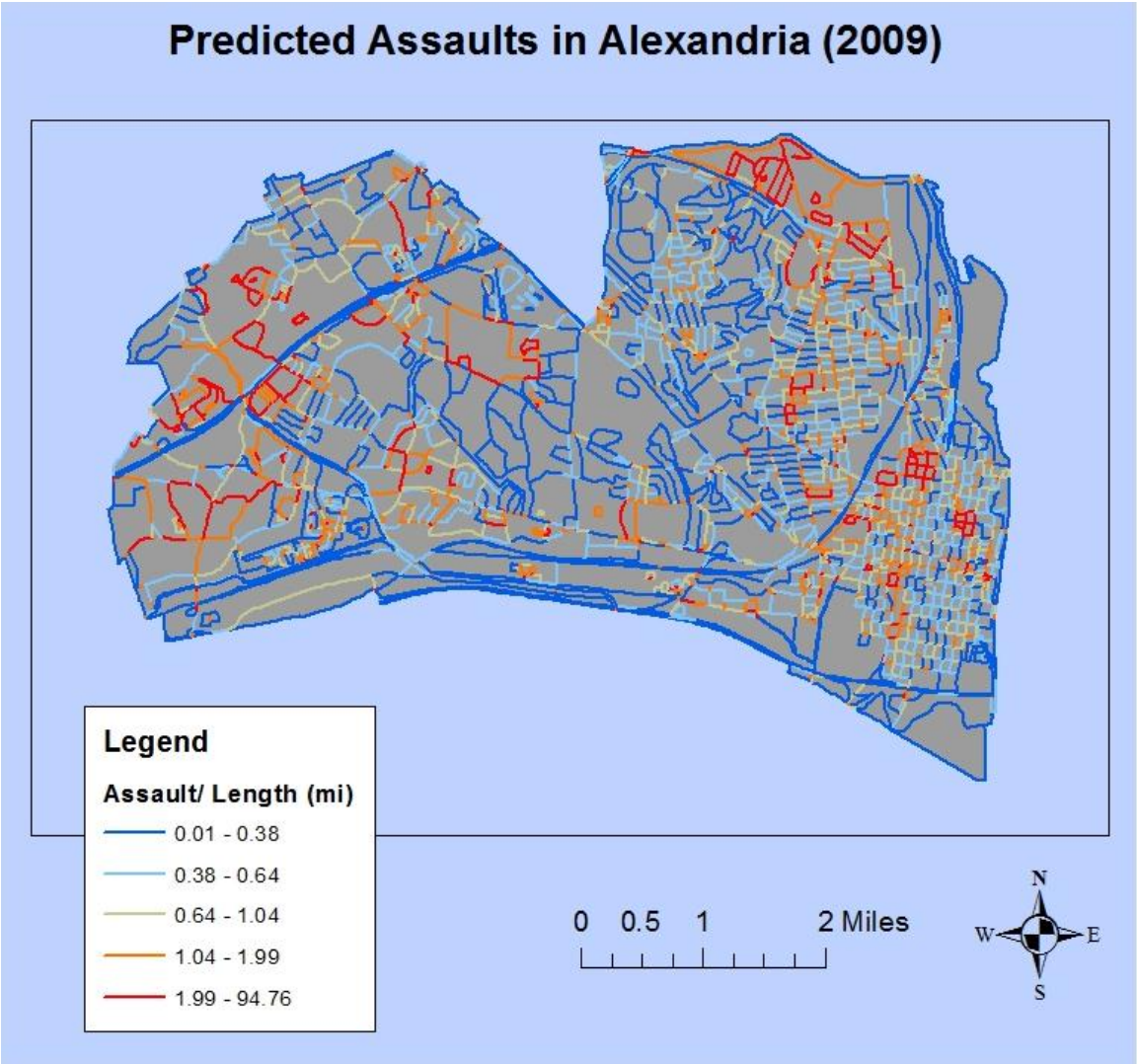

Figure 35: Predicted assault values over polylines for 2009 in Alexandria, VA.

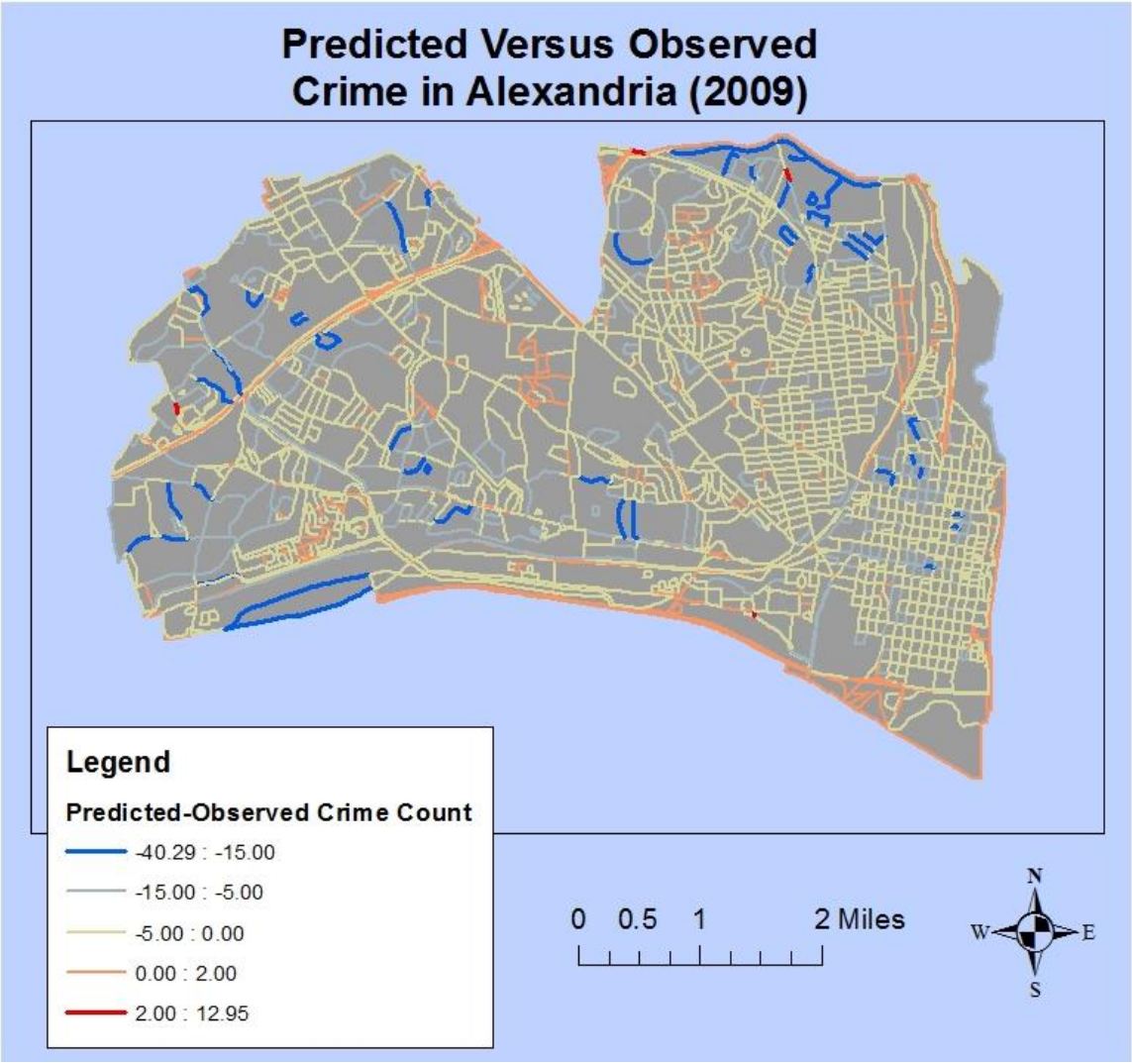

Figure 36: Predicted-Observed Crime Count

CHAPTER 6. VARIABLE SELECTION AND MULTIVARIATE VISUALIZATION

6.1 Variable Selection

While I use the models in the previous chapter to compare two data structures (area-based and road-based) and predict crime, these models are not the best to evaluate how important/significant each variable is in the assessment of crime, with multicollinearity being a big issue. I will now discuss two different methods that will resolve this problem and accurately select important variables. In statistics, variable selection is the process of selecting a subset of relevant variables for use in model construction. An assumption when selecting variables is that the data may contain many redundant or irrelevant variables that provide little information than a smaller subset of those variables would. Variable selection seeks to explain the data in the simplest way possible, without any unnecessary predictors that add noise to the estimation of other quantities.

I look at two methods of variable selection in this chapter. First I use random forest modeling, using conditional random forest variable importance to evaluate each variable conditional on every other variable in the data set. The second method I use finds the variables that are most highly correlated with crime and creates linear combinations of these variables using principal components analysis. Each of these methods provides different perspectives on the data, and address multicollinearity issues.

6.1 Random Forests and Variable Importance

A random forest is a popular tool used in classification and regression that grows many decision trees in order to appropriately classify objects and predict response values (Breiman, 2001). This tool has good predictive accuracy and easily accommodates many more variables than used in this dissertation. In order to build decision trees, the predictor space is divided into a number of regions and a prediction is made for a given observation based on the mean or mode of a set of training observations within the region containing the given observation. Using notation based on James et al. (2013), this means that for a set of values $X_1, X_2, \dots, X_n$, the predictor space is divided into j distinct non-overlapping regions $R_1, R_2, \dots, R_j$. For each observation in region R_j, the prediction is the mean of the response values for the training observations within R_j.

Decision trees are built using recursive binary splitting. This means that each branch of the tree is divided into two branches at each split of the predictor space. At each step, the best split is decided at that current step rather that looking ahead to see which split would create the best tree in the future. For regression, select the predictor X_j and cutpoint s such that splitting the predictor space into the regions $R_1(j,s) = \{X|X_j < s\}$ and $R_2(j,s) = \{X|X_j \geq s\}$ leads to minimizing the residual sum of squares (RSS):

$$\sum_{i:x_i \epsilon R_1(j,s)} (y_i - \hat{y}_{R_1})^2 + \sum_{i:x_i \epsilon R_2(j,s)} (y_i - \hat{y}_{R_2})^2 \tag{31}$$

where $\hat{y}_{R_1}$ is the mean response for the training observations in region $R_1(j,s)$ and $\hat{y}_{R_2}$ is the mean response for the training observations in region $R_2(j,s)$ (James et al., 2013).

This process is repeated at the next level to further minimize the RSS, with one of the previously split regions being split further into two more pieces. This continues until a pre-defined stopping rule is reached.

Decision trees have high variance, with widely differing results depending on how the training data is compiled. In order to compensate for this, random forests uses bootstrapping, also known as bagging in the context of decision trees. Bagging takes repeated samples from the data set and average over all of the resulting predicted values. That is, decision trees are created for each data set b of B bootstrapped data sets, the predicted value $\hat{f}^b(x)$ is calculated, and then the average is taken of the predicted values over all B data sets:

$$\hat{f}_{bag}(x) = \frac{1}{B}\sum_{b=1}^{B}\hat{f}^b(x). \tag{32}$$

For each tree, the data set that is created by sampling with replacement leaves out about one-third of the data. The observations that are left out of the sample are known as the out-of-bag (OOB) observations. A prediction can be made for the i^{th} observation using all of the trees for which that observation is OOB.

As opposed to simple decision trees, when producing a random forest, for each split, the algorithm assesses a random sample m of p variables as split candidates and picks the best split using the best candidate variable. A new sample of m variables is taken at each split. For regression, the number of variables considered at each split is usually taken to be equal to $p/3$, where p is the total number of predictor variables.

A measure of variable importance is the value of how much the RSS is decreased from splits over a given predictor variable, average over all B bootstrapped trees. The

first of variable importance is the mean decrease of accuracy on the OOB samples when a certain variable is excluded from the model. If this value is large, then the variable will be considered more important. A second measure is the total decrease in node impurity, resulting from splits over a given variable, averaged over all trees. Node impurity is measured by the RSS from the training data set. Similarly, a larger value indicates a more important variable.

The variable importance measures described above can be used to select which predictor variables are the most relevant to the response variable. However, issues can arise when many of the predictor variables are highly correlated (Strobl et al., 2008). This is due to the fact that typical variable importance measures are measures of marginal importance, whereas in the case of highly correlated variables the conditional effect of each variable may be more appropriate. A variable that may appear to be influential may actually be entirely independent of the response when it is considered from the perspective of being conditional on another variable with which it is highly correlated.

Strobl et al. (2008) develops a different variable importance measure using a conditional permutation method, which they show to be a more reliable measure in showing the true impact of each variable. Instead of the simple permutation of variable *m* given in the typical variable importance measure, the conditional method is carried out by having variable *m* being permuted within $Z = z$, where Z is the group of all other variables in the data set. This will preserve the correlation structure between *m* and all of the other predictor variables. If variable *m* and Z variables are independent, both methods will yield the same results. However, if the two are correlated, the original

variable importance assessment will increase the importance of correlated predictor variables.

Formally, there are four steps in calculating conditional variable importance. The first step is to compute the oob (out-of-bag) prediction accuracy before permutation:

$$\frac{\sum_{i\epsilon\bar{B}^{(t)}} I(y_i = \hat{y}_i^{(t)})}{|\bar{B}^{(t)}|} \tag{33}$$

where y_i is the response for observation i, $\hat{y}_i^{(t)}$ is the predicted class for observation i, and $\bar{B}^{(t)}$ is the oob sample for tree t with $t\epsilon\{1, \dots, ntree\}$. Second, for each variable Z to be conditioned on, cutpoints are extracted that split the variable in the current tree and a grid is created by dividing the sample space at each cutpoint. Then within the grid, the values of variable X_j are permuted and then the post-permutation oob prediction accuracy is calculated:

$$\frac{\sum_{i\epsilon\bar{B}^{(t)}} I(y_i = \hat{y}_{i,\pi_j|Z}^{(t)})}{|\bar{B}^{(t)}|} \tag{34}$$

where $\hat{y}_{i,\pi_j|Z}^{(t)}$ is the predicted class for observation i after permuting its value of variable X_j within the grid. Finally, the difference between the prediction accuracy before and after permuting gives the variable importance of X_j for one tree, which is then averaged over all trees (Strobl et al., 2008).

In **R** (R Core Team, 2013) the '**randomForest'** package supports computing the random forest estimates and variable importance. The **importance()** function in this package yields two variable importance measures. The R '**party'** package and its

function **cforest()** supports computing conditional random forests and their associated variable importance measures.

6.2 Condition Random Forest Results

6.2.1 Alexandria Crime and Assaults

I obtain variable importance measures from the conditional random forests for Alexandria in order to identify more accurately the most important crime-related variables and which variables will be the most valuable when visualization is restricted to using two explanatory variables. One weakness of using the conditional variable importance measurements is the computation time. The number of variables and observations significantly adds to the overall run time. To compensate for this, bootstrap samples were taken of the data and the variable importance values were measured over all of the samples. Even with these samples, the full crime data set took over 24 hours to run.

Figures 37 and 38 show the conditional random forest variable importance plots for all crime and assaults, respectively. Variables with values farther from zero are more important.

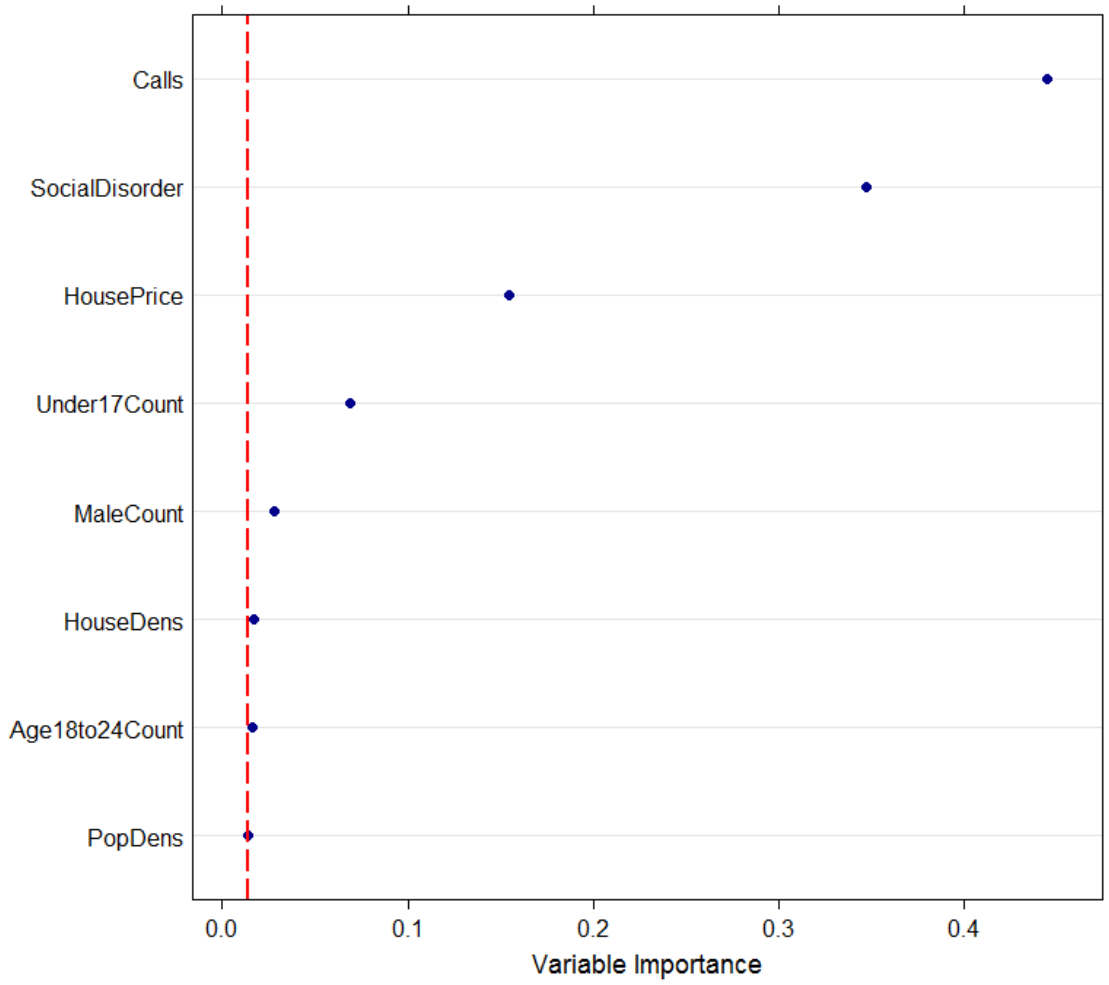

Figure 37: Conditional random forest for the full Alexandria crime data set (Variables to the right of the dashed line are significant).

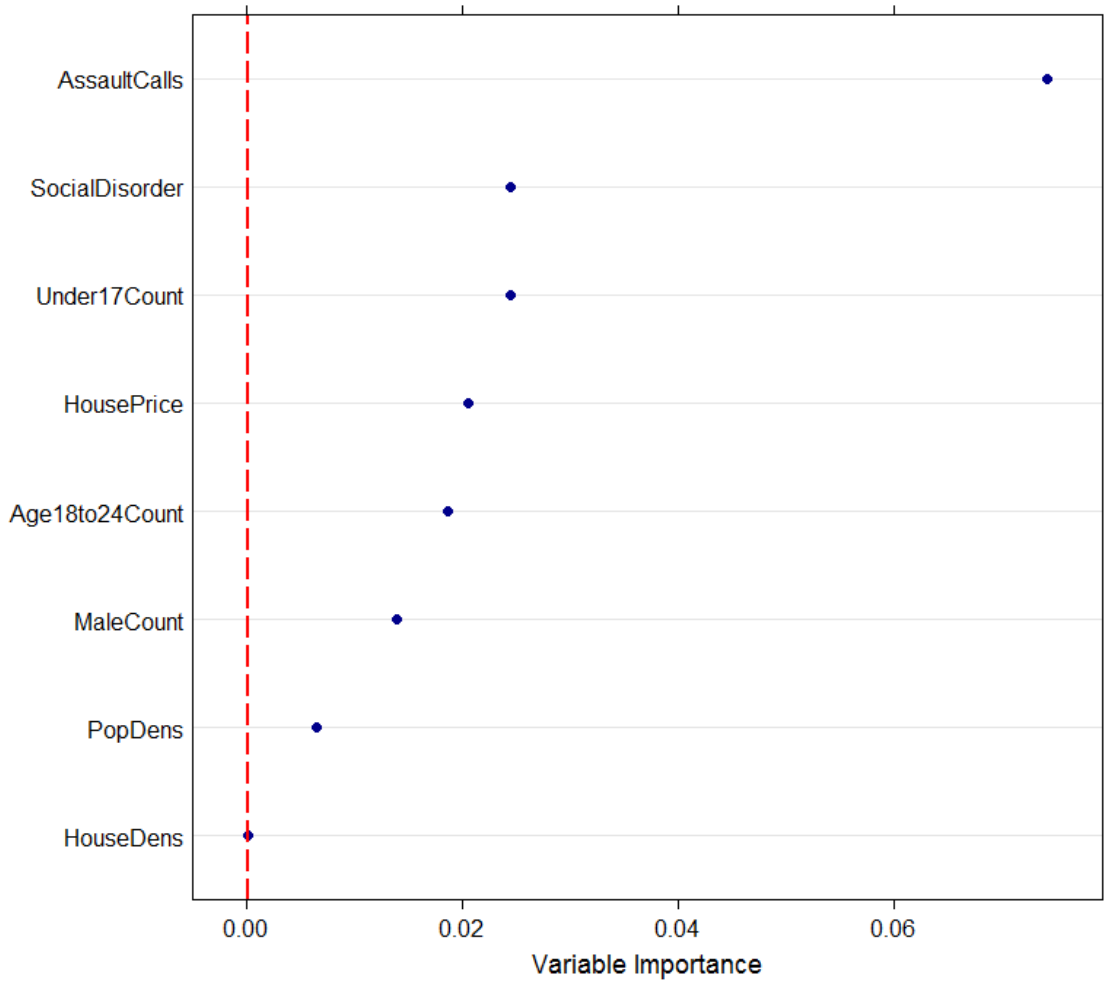

Figure 38: Conditional random forest for the Alexandria assault data set (Variables to the right of the dashed line are significant).

For the full crime data set, the conditional variable importance gives Service Calls, Social Disorder and Housing Price as the most important variables. Anything on or to the left of the red line is not important in the model. Housing density, population density, and the number of 18 to 24 year olds are not significant in this conditional random forest model. For assaults, the variables Assault Calls, Social Disorder, and Under 17 are the most important. Notice that the count of males and 18 to 24 years olds

are more important in this model than in the full crime data set, while housing price is not as significant.

When using these variables, the conditional random forest models do not fit as well as the CAR models as shown in Table 22. Note that while the random forest modeling takes into account the multicollinearity issue, it does not incorporate any spatial component, which may affect the model fits.

Table 22: Random Forest Mean Squared Error

MODEL	MEAN SQUARED ERROR
RANDOM FOREST: ALL CRIME	12.25
CAR MODEL: ALL CRIME	1.10
RANDOM FOREST: ASSAULTS	1.67
CAR MODEL: ASSAULTS	0.33

6.2.2 San Francisco Crimes

Figure 39 shows the conditional random forest variable importance plot for San Francisco, CA. Here, Elevation, Male Count and Housing Density pop up as the most importance in the conditional variable importance, and thus are most closely related to the criminal activity. Speed Limit is not at all a significant variable in the data set.

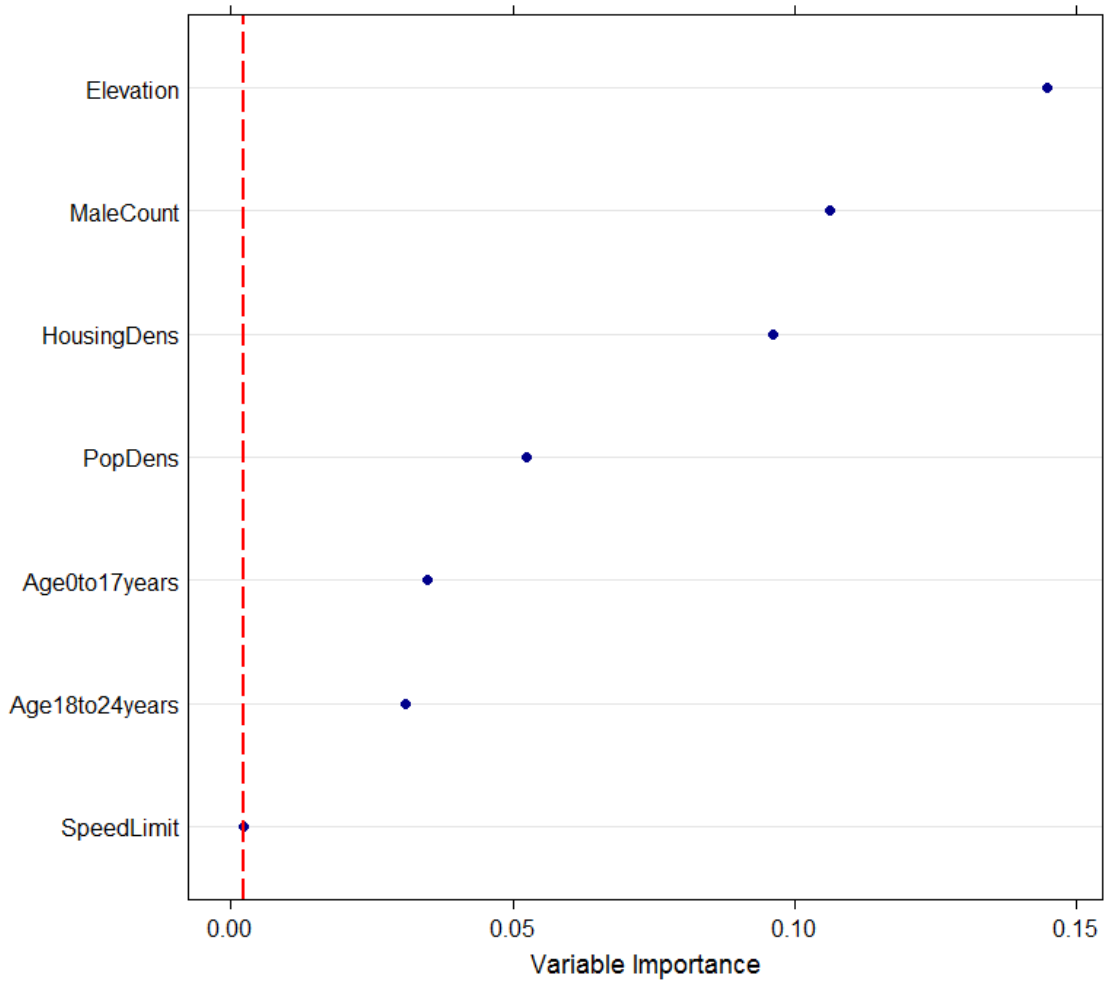

Figure 39: Conditional random forest for the San Francisco crime data set (Variables to the right of the dashed line are significant).

6.3 Supervised Principal Components Analysis

Rather than looking at each variable one at a time to assess variable importance, linearly independent combinations of these variables can be used. Principal Components Analysis (PCA) is a dimension reduction tool that can reduce the number of variables into a smaller set that still contains a majority of the information from the original set. This resolves the issue of multicollinearity in the data set. It uses linear transformations

of possibly correlated variables to create a sometimes smaller set of linearly independent, uncorrelated variables, known as principal components. The first principal component will account for the largest amount of variability among the variables, with each of the following uncorrelated components having the next highest variance.

Formally, let $\boldsymbol{X}$ denote the matrix of data with the j variables making up the columns of the matrix and the i observations making up the rows of the matrix. $\boldsymbol{X}$ can be divided up into three parts as follows:

$$\boldsymbol{X} = \boldsymbol{P}\boldsymbol{\Delta}\boldsymbol{Q}^{\boldsymbol{T}} \tag{35}$$

where $\boldsymbol{P}$ is a matrix of left singular vectors, $\boldsymbol{Q}$ is a matrix of right singular vectors, and $\boldsymbol{\Delta}$ is the diagonal matrix of singular values (Hervé and Williams, 2010). The matrix $\boldsymbol{Q}$ gives the coefficients of the linear combinations used to compute the factor scores. The matrix of principal component factor scores $\boldsymbol{F}$ is obtained as follows:

$$\boldsymbol{F} = \boldsymbol{P}\boldsymbol{\Delta} = \boldsymbol{X}\boldsymbol{Q}. \tag{36}$$

I selected the variables to use in principal components following a supervised principal component approach developed by A. Vidyashankar (2014). His approach screens variables in possibly high dimensional data sets for use in principal components in order to help address computation and interpretation issues. The procedure uses

multiple 50% bootstrap samples of the cases to obtain the correlation distributions between the dependent variable with each of the candidate regression variables.

Note that the distribution of correlations for a variable may show interesting patterns related to the subsets of cases sampled and begin to suggest localized variable importance. Local variable importance at the case level can also be addressed using random forests. This could be important in crime studies, and I leave the study of this a topic for future research.

The more supervised principal components procedure uses a chosen percentile of the resulting correlation distribution and the cutoff threshold for variable selection. I chose to select variables whose 70^{th} percentile was a positive correlation above .3 for the Alexandria data and above .4 for the San Francisco data. The 30^{th} percentile could be used to select variables based on negative correlations as well. I run principal components analysis on these variables. For the full crime data, the five selected variables include the two age variables (count of those under 17 and between 18 to 24), the number of males, the calls for service, and social disorder. A table of these correlations is given below. The selection for the assault data set is equivalent except we replace calls for service with assault-related calls for service.

Table 23: 70^{th} Percentile Correlation of Crime and Crime-Related Variables for Alexandria, VA

Variables	**Crime Correlations**
Under 17 Count	0.328
Age 18 to 24 Count	0.323

Male Count	0.322
Population Density	0.263
Housing Density	0.236
House Price	-0.117
Calls	0.448
Social Disorder	0.455

I computed principal components analysis using the **R** function princomp(). Running the principal components analysis for the Alexandria data set gives loadings in Table 24. The first principal component is roughly the average of all of the standardized variables, with slightly higher correlations for young males (that is, the two young age categories and males). The second principal component a contrast between the social disorder and calls for service and the counts of the 0 to 17 and 18 to 24. This suggests there are a greater number of calls relative to the number to the population values. An alternative variable thus might be a calls for service to young male population ratio.

The plot following the principal components table gives an illustration of the proportion of variance covered. The first two components retain over 90% of the original variability of the data set, suggesting that these are the two important principal components which will be very helpful in models and graphs. The assault data set will give similar results, but with the full calls for service data set replaced with assault calls for service.

Table 24: Principal Components Correlations for the Full Crime Data Set for Alexandria, VA

Variable	Component 1	Component 2	Component 3	Component 4	Component 5
Social Disorder	-0.294	0.648	0.532	-0.456	
Calls	-0.314	0.629	-0.522	0.480	
Under 17 Count	-0.511	-0.259	0.553	0.605	
Age 18 to 24 Count	-0.527	-0.243	-0.290	-0.266	-0.713
Male Count	-0.526	-0.243	-0.233	-0.352	0.697

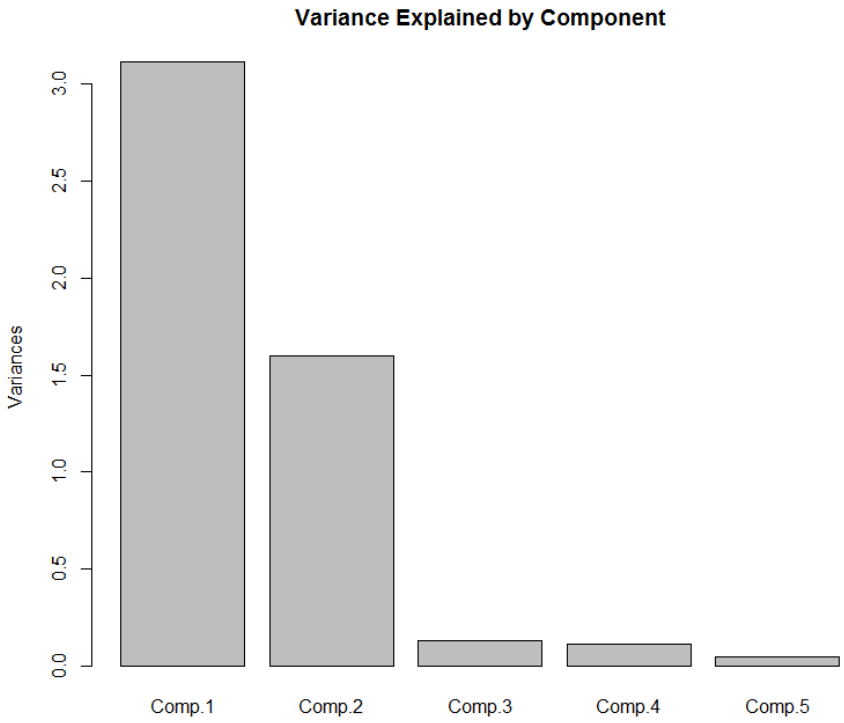

Figure 40: Bar Chart of Variance Explained by Each Principal Component

I ran both the zero-inflated negative binomial model and the CAR model with just these two principal components to show that the model-fitting measurements will give results as using all of the variables. Table 25 shows the mean squared error, AIC, and DIC values. Using only these two variables is comparable to using the full set of variables in Table 19. A big advantage of using principal components analysis resolves the issue of multicollinearity and creates a simplified, reduced data set without too much loss of information.

Table 25: Mean Squared Error, AIC and DIC using Principal Components

	Full Crime Zero-Inflated	**Assault Zero-Inflated**	**Full Crime CAR Model**	**Assault CAR Model**
Mean Squared Error	1.30	1.48	0.94	0.28
AIC	14959	6780	NA	NA
DIC	NA	NA	13254	6028

The same procedure is applicable to the available crime-related variables in San Francisco, CA. Here the highest correlated variables are Housing Density, Population Density, Male Count, and Age 18 to 24 Count. These are the four variables that I will select to be in the principal components analysis.

Table 26: 70th Percentile Correlation of Crime and Crime-Related Variables for San Francisco, CA

Variables	**Crime Correlations**
Under 17 Count	0.386
Age 18 to 24 Count	0.423
Male Count	0.520
Population Density	0.527
Housing Density	0.540
Elevation	-0.137
Speed Limit	0.095

Table 27 shows the results of the principal components analysis. Once again, over 90% of the variance from the original data set is explained with the first two principal components, so I focus on those specifically. Based on the correlations, the first component seems to be a mixture of population density, housing density, and male count. The second component is a contrast between housing and population densities and Male and Age 18 to 24 counts with the age groups weighted most heavily.

Table 27: Principal Components for the Full Crime Data Set for San Francisco, CA

Variable	Component 1	Component 2	Component 3	Component 4
Housing Density	-0.545	0.366	0.419	0.627
Population Density	-0.570	0.295	0.131	-0.755
Male Count	-0.544	-0.225	-0.787	0.186
Age 18 to 24 Count	-0.287	-0.853	0.433	

I use the multivariate visualization tool DPnet in the next section for Alexandria, VA. I show both the most important variables from the conditional random forests and the linear combinations of variables given from the principal components analysis to see the patterns these variables have with regards to crime. The principal components used may be difficult to interpret, as they are linear combinations of several variables; however, they may also lead to a new way to conceptualize crime and how it relates to multiple variables at a time.

6.3 DPnet Results

I use DPnet to partition crime counts for road segments in conjunction with two additional variables. First I focus on the variables given in the conditional random forest modeling, followed by the variables created in principal components analysis. Dynamically partitioned maps will draw polylines to represent road segments. The crime counts slider at the top has two thresholds that distinguish low, middle and high values, shown in blue, gray and red, respectively. Below the slider are thresholds values. The sliders at the right and bottom axes can be changed using the chosen covariates I would like to analyze. The average crime count of road segments highlighted in a panel appears at the top right of each panel. If there is a change in segment crime rates based on the slider variables, this will be reflected in the pattern of counts. The mean counts for the road segments highlighted in a panel are used as fitted values in a simple model, and the R-squared describes the quality of the fit at the lower right of the plots.

Figure 41 is a snapshot of what DPnet can do with the full crime data set over road networks with two other variables. Here I used two of the most important variables, property values and police calls for service, mapped against the fractional crime counts on the road segments. The three variable groups show patterns that are almost impossible to notice when looking at the variables one at a time. For example, Figure 42 gives a close-up of the upper left corner of this plot. This shows a cluster of segments in blue, signifying a low crime count, in an area with high housing prices and low number of calls, which is what might be expected.

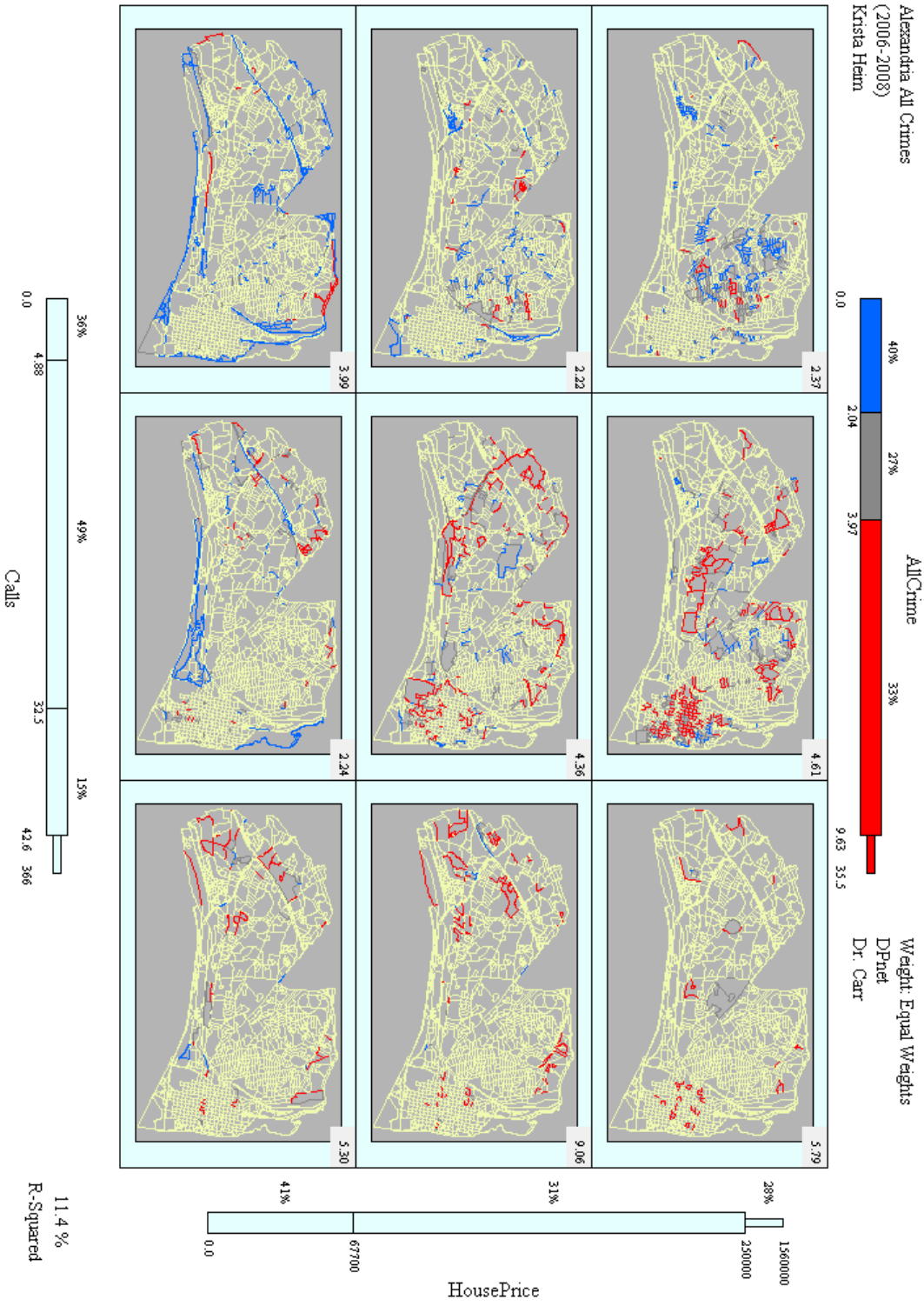

Figure 41: DPnet for all crimes with covariates police calls for service and house property sales.

Figure 42: Zoom in on the upper-left corner of DPnet for All Crimes.

Figure 43 gives the DPnet of the smoothed assault counts with the two most important variables chosen from the conditional random forest, the social disorder variable and the number of people under 17. There is a trend here with high crime along the same roads as high number of social disorder calls and a high count of people under 17 years of age, and low number of crimes along the roads with low amount of social disorder and number of those under 17. The top right panel shows a cluster of roads with high crime rates. All of the right panels with highlighted roads associated with high social disorder have high crime rates. Figure 44 zooms in on the middle left panel

associated low social disorder and middle amounts of people under 17. You can see an entire section of blue segments here, representing low crime.

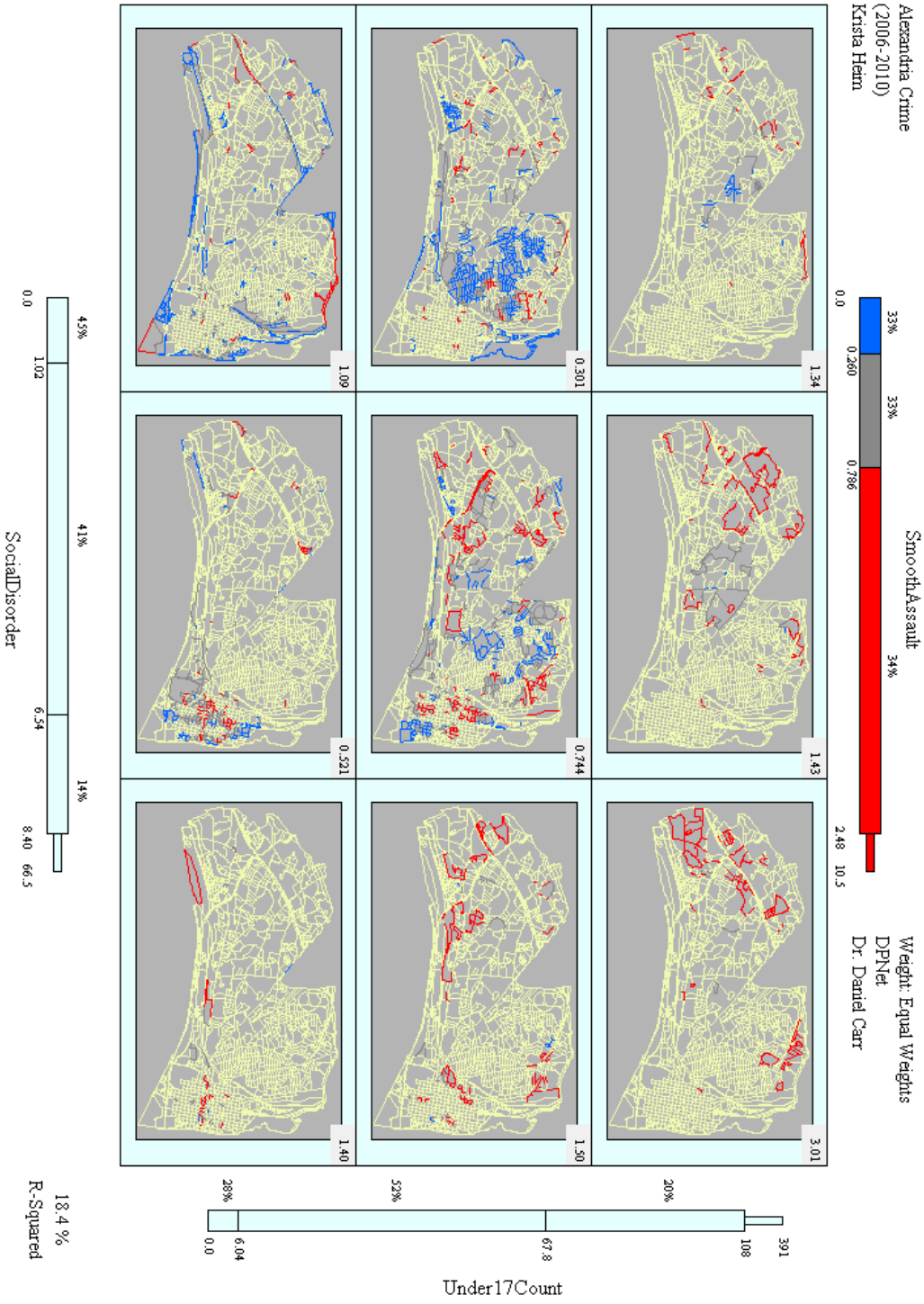

Figure 43: DPnet for Assaults with covariates Social Disorder and Under 17 Counts.

Figure 44: Zoom in on middle-left section of DPnet for Assaults.

So far, the partition has produced small R^2 values, indicating the panels don't fit the road segment crime counts very well. Now I will map the two variables created in principal components analysis against the crime. I described the first principal component as "Average of Calls, Social Disorder, and Young Male Population", since it represented linear combination of these variables. The second component I will define as "Calls/ Social Disorder and Young Male Population Contrast", as this is mostly a combination of the calls for service and social disorder variables. Figure 45 gives the full DPnet view. The middle-right section includes one very large crime value that seems to drive it to be in its own category. I do not know what is occurring at the segment that is

making the crime counts high at that specific location, but as stated before there is always the possibility of geocoding errors. The R^2 value has increased to over 66%. Following the full map are two snapshots in Figures 46 and 47. The upper-middle section, representing large "Calls/ Social Disorder and Young Male Population Contrast" values and middle-sized "Average of Calls, Social Disorder, and Young Male Population" values, contains the large cluster of high crime in Old Town Alexandria. The lower-middle section, representing small "Calls/ Social Disorder and Young Male Population Contrast" values and middle-sized "Average of Calls, Social Disorder, and Young Male Population" values, has the cluster of low crime just northwest of the downtown Alexandria area. Although these components may be hard to interpret, as they are no longer count values and have negative and positive numbers, they show definite patterns with the crimes that the original variables by themselves did not give.

Figure 48 gives a similar DPnet, but removes those high segments that were in the middle-right section. This helps us see some patterns that may have been masked by these high values. For example, the upper-right section has a few very high segments, with average crime count of 18 in that section. There is a low patch of crime in the middle-left section, representing middle values of "Calls/ Social Disorder and Young Male Population Contrast" and low values of "Average of Calls, Social Disorder, and Young Male Population".

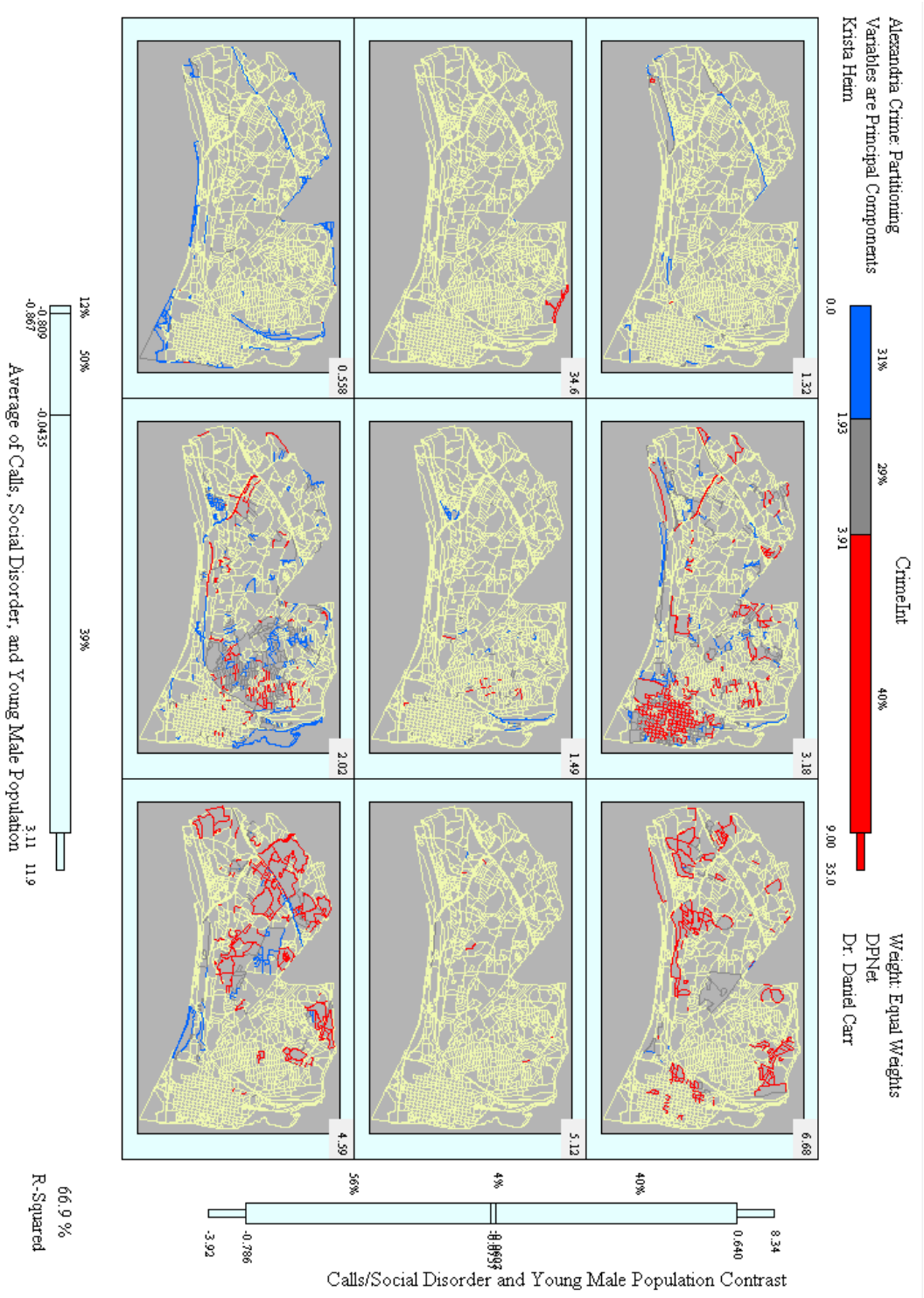

Figure 45: DPnet of the crime counts compared with the first two principal components.

Figure 46: Upper-middle section of DPnet of the crime counts compared with the first two principal components

Figure 47: Lower-middle section of DPnet of the crime counts compared with the first two principal components

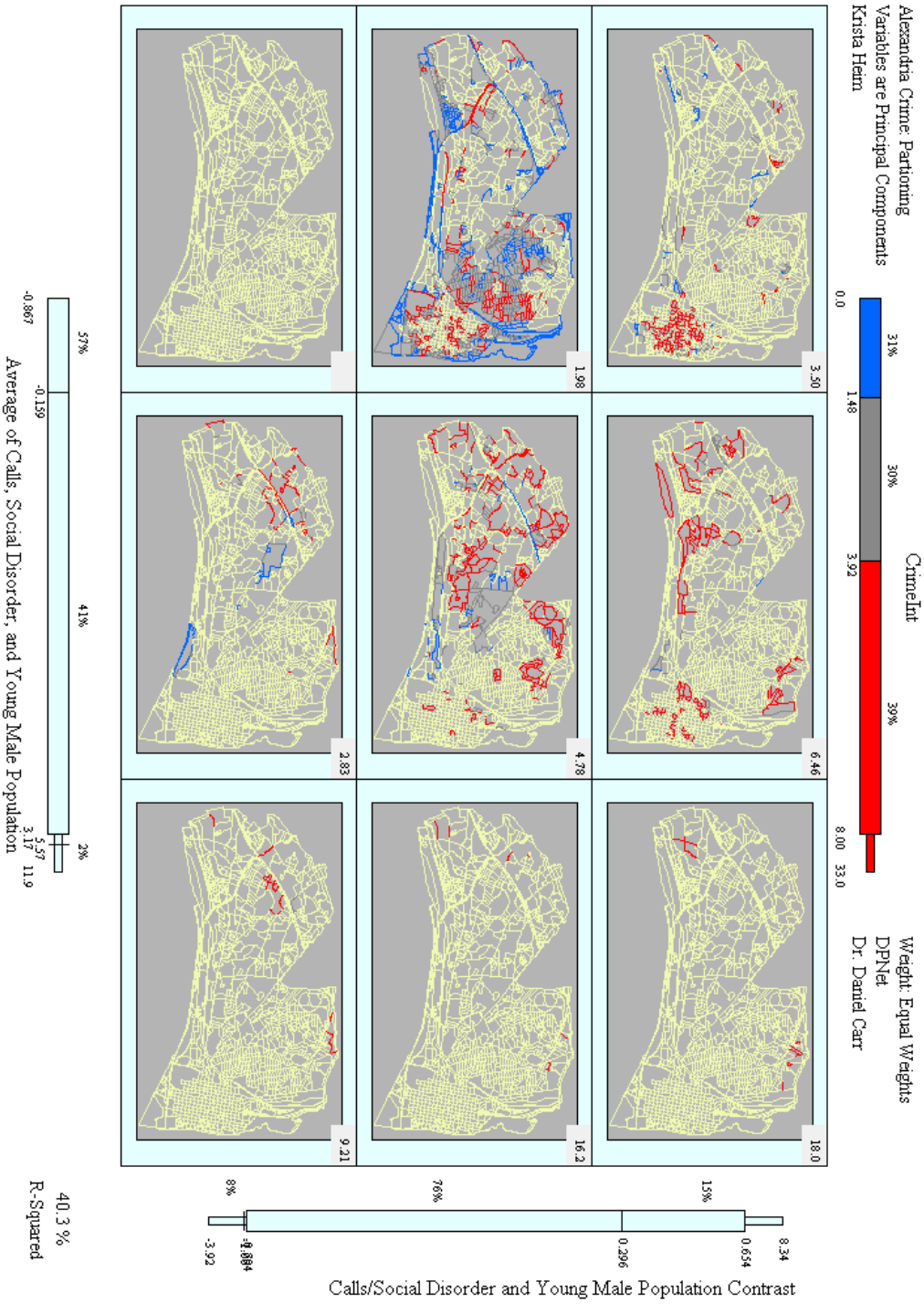

Figure 48: DPnet of the smoothed crime counts compared with the first two principal components, ignoring high crime segments.

CHAPTER 7. CONCLUSIONS AND FUTURE WORK

In this research, I developed and adapted methods to support the modeling and visualization of crime data and covariates indexed by road segments. This is part of a broader vision that seeks to put analysis of data indexed by lines and polylines on a more equal footing with analysis of data indexed by points or polygons. Methods that convert point and area data to line-indexed data enable line-indexed data to be used as a unifying framework for modeling and visualization. In the line-indexed framework common general concepts such as neighbors, distance, and spatial correlation remain relevant, but the exact usage can be adapted to the framework as motivated by phenomena such as crime.

The choice of crime data is strategic. The theory of the criminology of place has focused attention on "micro" analysis where road segments become particularly relevant. The clusters of road segments in Weisburd et al. (2012) and the graphics reveal a variety of crime patterns on roads that can be obscured by area-based analysis. The extensive criminology of place literature is persuasive and motivates my analysis along street segments. While many variables chosen here aren't directly related to the criminology of place, they still call attention to the benefits of crime analysis along this spatial unit and support taking next steps in using a wider variety of models and graphics to look deeper into crime patterns with the limitation of readily available data and its quality. The

models found most of the variables suggested to be related to crime indexed by roads. Places where predictions are poor perhaps can be improved by using more specifics of the street segments and incorporating more details of place.

I assigned fractional counts of crime to surrounding road segments based on inverse distance weighting. I created a unique smoothing algorithm in R that reweights crime counts over road segments according to the distances to nearest segments and the angle at which they meet. This results in a smooth visualization of crime over roads to be more easily interpreted by law enforcement by reducing noise to help see the patterns of crime more clearly. The smoothing algorithm was used for Alexandria, VA and San Francisco, CA, with visualizations in ArcMap.

Both point and area data (such as Census block data) were converted into polyline statistics, providing a unified framework for modeling along street segments. I focused on several models that represented different facets of the data, including the zero-inflated negative binomial model and the Poisson-Gamma CAR model. I compared models that aggregate data to areal units with models using counts along road segments. The road-based CAR model used uniquely-defined matrices depending on nearest road segments and distance between those segments. In the example of Alexandria, VA, the road-based models gave better fit results and could better predict crime (and assaults) than when they were aggregated over area units. This unique method gives very accurate prediction results at a local, "micro" level and is a step forward in predictive policing.

The interpretation of crime covariates in the models is complicated by their correlations. To partially address this, I took some early steps that analyze the model covariates in terms of their importance in the models and also used supervised principal complements to produce variables that were independent. For the full crime data set of Alexandria, VA, the most important variables from the random forests included the calls for service, social disorder, and housing prices. For the assault data set, the most important variables were found to be the assault calls for service, social disorder, and the number of people under 17 years of age. It is interesting to note that social disorder (complaints, noise violations, etc.) is highly important in both cases, while age is more relevant to the assault crimes. Elevation, count of males, and housing density were found to be the top three most important variables for San Francisco, CA. This supports the idea that different types of variables can describe crime in different locations. Principal components comprised of linear combinations of highly correlated variables were generated using principal components analysis. Models using the first two components were competitive to models using the full set of variables. Such work may be helpful in terms of variable selection and possibly in terms of the evolution of the criminology of place theory.

Using the most important variables from conditional random forests and the two principal component variables, DPnet created a multivariate visualization along the street segments in order to explore patterns visually that aren't typically possible to explore. This visualization was used only for Alexandria, VA. The software is not yet able to handle the number of road segments in San Francisco, CA to be able to create a DPnet

example for this location. The principal components may be difficult to interpret, but provide a very strong pattern visually and would be interesting to develop further.

There are a number of different things I could do to extend my research. I would like to explore more rigorously the idea of using a CAR model of polylines rather than polygons. I could look more closely to identify special features of those road segments that are poorly fit by the models. Alternative assessments of modeling and variable importance could be used. Structural equation modeling (SEM) could be used, which constructs latent variables (variables that are not measured directly but may still have effects on the data) in order to capture the unreliability of measurements in the previous models. I could also explore a number of different projection and weighting schemes for moving crime to road segments and for smoothing those points for visualization. I could separate crimes that tend to be closer to intersections and crimes that tend to be closer to midpoints and model separately. I could consider a weighting that incorporates the direction of the flow of traffic on a street. Whether or not the traffic flow is on a one-way or two-way road may make a difference in the analysis of crimes on these segments. Using fractional counts might not be easily interpretable for the police force. It might make sense to convert back to simple counts.

I would like to have more data to explore, as data at the micro level can be hard to come by. In general, the utility of models and graphics depends on the availability and quality of the relevant variables. The "criminology of place" studies suggest that local features can make a big difference in crime rates. Often data is not gathered. When local data is gathered there are often barriers to obtaining data and work to do in preparing the

data for analysis. Many of the variables that the Census provides are only at the block group level, which covers a much larger area than the block data. This is much less useful to look at when focusing on crime at the local level. Some of the variables included in this data set are race, poverty status, and single family home, which would be incredibly useful as crime predictors if they were given at a finer scale. With the data I do have, it would be useful to explore smaller levels of the age category, to focus in more detail on teens/adolescents rather than having the 0-17 and 18-24 variables. Having an adolescent age category could better be a better representation of the group of criminal offenders. Other variables that may be useful in the future with the increase of availability of data and technological advances include a cell phone-based assessment of street traffic, more readily available data on vacant homes, building uses, and assessments of pedestrian traffic on pathways. It may also be useful to incorporate what local people and law enforcement know about the areas into the analysis in some way.

This type of modeling/visualization could be extended to other types of data aside from crime. My model could in particular be useful for looking at car accidents, as the majority of the time these occur on roads. The modeling and visualization methods given here are flexible for handling many different types of data and covariates.

REFERENCES

Abdi, Hervé and Lynne J. Williams (2010). Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*. John Wiley & Sons, Inc.

Anselin, Lucien (1995). Local Indicators of Spatial Association- LISA. *Geographical Analysis 27(2)*.

Baddeley, Adrian and Rolf Turner (2005). spatstat: An R Package for Analyzing Spatial Point Patterns. *Journal of Statistical Software 12(6),* 1-42.

Besag, Julian (1974). Spatial Interaction and the Statistical Analysis of Lattice Systems. *Journal of the Royal Statistical Society 36 (2),* 192-236.

Beavon, Brantingham and Brantingham (1994). The Influence of Street Networks on the Patterning of Property Offenses (Ronald V. Clarke, Ed.). In *Crime Prevention Studies,* Volume 2, 115-147. Willow Tree Press, Inc.

Brantingham and Brantingham (1995). Criminality of place. *European Journal on Criminal Policy and Research 3(3)*, 5-26.

Breiman, Leo (2001). Random Forests. *Machine Learning*. 5-32.

Cameron, Colin A. and Pravin K. Trivedi (1998). *Regression Analysis of Count Data*. Cambridge University Press.

Campenella, Richard (2002). *Time and Place in New Orleans: Past Geographies in the Present Day.* Pelican Publishing.

Carr, Daniel B., John F. Wallin, and D. Andrew Carr (2000). Two New Templates for Epidemiology Applications: Linked Micromap Plots and Conditioned Choropleth Maps, *Statistics in Medicine 19 (17–8),* 2521–2538.

Carr, Daniel B., Linda Williams Pickle (2010). *Visualizing Data Patterns with Micromap*s. Chapman & Hall/CRC Interdisciplinary Statistics.

Carr, Daniel B., Denis White and Alan M MacEachren (2005). Conditioned Chloropleth Maps and Hypothesis Generation. *Annals of the Association of American Geographers, 95(1)*, 32–53.

City and County of San Francisco (2014). *San Francisco Open Data Portal*. <data.sfgov.org>.

Cleveland, William S. and Robert McGill (1984). Graphical Perception: Theory, Experimentation, and Application to the Development of Graphical Methods. *Journal of the American Statistical Association, 79(387)*, 531-554.

Cohen, Lawrence and Marcus Felson (1979). Social Change and Crime Rate Trends: A Routine Activity Approach. *American Sociological Review*, Volume 44, 588-608.

Cressie, Noel A. C. (1993). *Statistics for Spatial Data*. New York: John Wiley & Sons.

Curriero, Frank C. (2006). On the Use of Non-Euclidean Distance Measures in Geostatistics. *Mathematical Geology 38(8)*, 907-926.

de Oliveira, Daniel, Daniel Neill, James Garrett Jr. and Lucio Soibelman (2011). Detection of Patterns in Water Distribution Pipe Breakage Using Spatial Scan Statistics for Point Events in a Physical Network. *Journal of Computing in Civil Engineering*. January/February 2011.

Eck, John E., Spencer Chainey, James G. Cameron, Michael Leitner and Ronald E.Wilson (2005). Mapping Crime: Understanding Hot Spots. *NIJ Special Report* (August 2005).

ESRI (2011). ArcGIS Desktop: Release 10. Redlands, CA: Environmental Systems Research Institute.

Faraway, Julian J. (2006). *Extending the Linear Model with R: Generalized Linear, Mixed Effects and Nonparametric Regression Models*. Chapman & Hall/CRC.

Gelman, Andrew, John B. Carlin, Hal S Stern and Donald B. Rubin (1995). *Bayesian Data Analysis*. Chapman & Hall/CRC.

Groff, Elizabeth R., David Weisburd and Sue-Ming Yang (2010). Is it Important to Examine Trends at the "Micro" Level?: A Longitudinal Analysis of Street to Street Variability in Crime Trajectories. *Journal of Quantitative Criminology*, Volume 26, 7-32.

Haining, Robert (2003). Spatial Data Analysis: Theory and Practice. Cambridge University Press.

James, Gareth, Daniela Witten, Trevor Hastie and Robert Tibshirani (2013). *An Introduction to Statistical Learning with Applications in R*. Springer Science + Business Media New York.

Kulldorff, Martin (1997). A Spatial Scan Statistic. *Communications in Statistics - Theory and Methods 26(6)*.

Kulldorff, Martin and Information Management Services, Inc. (2009). SaTScan v8.0: Software for the spatial and space-time scan statistics.

Lee, Duncan (2013). CARBayes: An R Package for Bayesian Spatial Modeling with Conditional Autoregressive Priors. *Journal of Statistical Software 55(13)*, 1-24.

Lee, Duncan and R. Mitchell (2012). Boundary Detection in Disease Mapping Studies. *Biostatistics 13(3)*, 415–426.

Levine, Ned (2006). Crime Mapping and the Crimestat Program. *Geographical Analysis*, Volume 38, 41-56.

Levine, Ned (2013). CrimeStat: A Spatial Statistics Program for the Analysis of Crime Incident Locations (v 4.0). Ned Levine & Associates, Houston, Texas, and the National Institute of Justice, Washington, D.C. June.

Liu, Hua and Donald E. Brown (2003). Criminal incident prediction using a point-pattern-based density model. *International Journal of Forecasting*, Volume 19, 603-622.

Miaou, Shaw-Pin, Joon Jin Song and Bani K. Mallick (2003). Roadway traffic crash mapping: a space-time modeling approach, *Journal of Transportation & Statistics 6 (1)*. 33-58.

Nagin, Daniel S. (2005). *Group-based modeling of development*. Cambridge: Harvard University Press.

Nettler, G. (1978). *Explaining Crime*. 2nd ed. Montreal: McGraw-Hill.

Osgood, D. Wayne (2000). Poisson-Based Regression Analysis of Aggregate Crime Rates. *Journal of Quantitative Criminology* 16(1).

Peterson, Erin E., David M. Theobald and Jay M Ver Hoef (2007). Geostatistical Modelling on stream networks: developing valid covariance matrices based on hydrologic distance and stream flow. *Freshwater Biology*, Volume 52, 267-279.

Ratcliffe, Jerry (2013). What Is the Future… of Predictive Policing? *The magazine of the center for evidence-based crime policy, Spring 2014*. George Mason University.

R Core Team (2013). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0, <http://www.R-project.org/>.

Rosling, Hans (2010). "The Joy of Stats". Wingspan Productions for BBC. Director & Producer; Dan Hillman, Executive Producer: Archie Baron.

Schabenberger, Oliver and Carol A. Gotway (2005). *Statistical Methods for Spatial Data Analysis*. Chapman & Hall/CRC Press.

Sherman, Lawrence W., Patrick R. Gartin and Michael E. Buerger (1989). Hot Spots of Predatory Crime: Routine Activities and the Criminology of Place. *Criminology*, Volume 27.

Short, M. B., M. R. Dorsogna, V. B. Pasour, G. E. Tita, P. J. Brantingham, A. L. Bertozzi and L. B. Chayes (2008). A Statistical Model of Criminal Behavior. *Mathematical Models and Methods in Applied Sciences*, Volume 18, 1249-1267.

Song, Joon Jin, M. Ghosh, S. Miaou and B. Mallick (2006). Bayesian multivariate spatial models for roadway traffic crash mapping. *Journal of Multivariate Analysis 97(1)*, 246-273.

Stern, H., Noel Cressie (1999). Inference for Extremes in Disease Mapping. *Disease Mapping and Risk Assessment for Public Health*. John Wiley & Sons, 63–84.

Strobl, Carolin, Anne-Laure Boulesteix, Thomas Kneib, Thomas Augustin and Achim Zeileis (2008). Conditional variable importance for random forests. *BMC Bioinformatics*, 9:307.

Sun, D., Tsutakawa R. K. and Speckman, P. (1999) Posterior distribution of hierarchical models using CAR (1) distributions. *Biometrika*, Volume 86, 341-350.

U.S. Census Bureau (2013). *American FactFinder*. <http://factfinder2.census.gov>.

Van Patten, Isaac T., Jennifer McKeldin-Coner and Deana Cox (2009). A Microspatial Analysis of Robbery: Prospective Hot Spotting in a Small City. *Crime Mapping: A Journal of Research and Practice 1(1)*, 7-32.

Ver Hoef, Jay M., Erin Peterson and David Theobald (2006). Spatial statistical models that use flow and stream distance. *Environ Ecol Stat*, Volume 13, 449-464.

Ver Hoef, Jay M., Erin Peterson, D. Clifford and R Shah (2014). SSN: An R Package for Spatial Statistical Modeling on Stream Networks. *Journal of Statistical Software* 56 (3), 1-45.

Vidyashankar, A.N. (2014). Supervised Principal Components with possibly high dimensional covariates. Manuscript in preparation.

Vold, George B., Thomas J. Bernard, Jeffrey B. Snipes (2002). *Theoretical Criminology*. Fifth Edition. Oxford University Press, Inc.

Weisburd, David, Shawn Bushway, Cynthia Lum and Sue-Ming Yang (2004). Trajectories of Crime at Places: A Longitudinal Study of Street Segments in the City of Seattle. *Criminology 42(2)*, 283-322.

Weisburd, David, Elizabeth Groff and Nancy Morris (2011). Hot Spots of Juvenile Crime: Findings from Seattle, Juvenile Justice Bulletin, Office of Juvenile Justice and Delinquency Prevention.

Weisburd, David, Elizabeth Groff and Sue-Ming Yang (2012). *The Criminology of Place: Street Segments and Our Understanding of the Crime Problem*. Oxford University Press.

Wilson, James Q. and George L. Kelling (1982). "Broken Windows: The police and neighborhood safety", *The Atlantic*, March 1, 1982.

BIOGRAPHY

Krista Heim graduated from St. Hubert Catholic High School for Girls in 2005 in Philadelphia, PA. She then received her Bachelor of Science in Mathematics from Arcadia University in Glenside, PA in 2009. In 2011, she received her Masters of Science in Statistical Science from George Mason University.