ANALYSING AND VISUALISING AREAL CRIME DATA

A Case Study of Residential Burglary in San Francisco, USA

Dissertation supervised by:

Professor Ana Cristina Costa, Ph.D

Professor Jorge Mateu, Ph.D

Professor Edzer Pebesma, Ph.D

February 2012

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my supervisors, Professor Ana Cristina Costa, Professor Jorge Mateu, and Professor Edzer Pebesma, for all of their kind help and assistance throughout the process of producing this thesis.

For his guidance and encouragement during the thesis progress sessions, I would like to express my gratitude to Professor Marco Painho.

I am also grateful to the all teaching staff at Instituto Superior de Estatística e Gestão de Informação, Universidade Nova de Lisboa and Institute for Geoinformatics, Westfälische Wilhelms-Universität Münster for their support and for sharing their invaluable knowledge during the course of my Master's studies.

A special thanks to Dr. Pierre Goovaerts and the staff at BioMedware for all of their kind assistance with their SpaceStat software.

ANALYSING AND VISUALISING AREAL CRIME DATA

A Case Study of Residential Burglary in San Francisco, USA

ABSTRACT

Methods to visualise and analyse areal social data are limited. A traditional approach is Choropleth mapping. However, the rates on which these maps are based can be unreliable in sparsely populated areas, and there may be visual bias when areas are irregularly sized. Another common method is to perform point interpolation at the centroids of the areas. This approach may only be valid when areas are regularly shaped and small.

This thesis explores how Area-to-Area and Area-to-Point kriging can be applied to analysing and visualising residential burglary rates in San Francisco, United States. Results are compared to the traditional methods used to analyse areal data. Additionally, the study investigates burglary hotspots and the relationship between socio-economic variables and burglary in the study area by conducting spatial and non-spatial regression analyses.

The study concludes that Area-to-Area and Area-to-Point Poisson kriging methods may improve on existing approaches to interpolating areal crime data. The visualisation of areal data is improved through the smoothing of rates based on small denominators, and visual bias may be decreased by using Area-to-Point kriging. Using the kriging estimates of these techniques as inputs into hotspot and regression analyses provides a useful way in which to explore relationships at different scales. However, caution should be exercised when utilising these methods due to their limitations.

KEYWORDS

Areal crime data

Area-to-Area Kriging

Area-to-Point Kriging

Deconvolution

Geographically Weighted Regression

Local Indicators of Spatial Association

Poisson Kriging

Regression

Residential Burglary

Residential Burglary Hotspots

San Francisco

ACRONYMS

ACS – American Community Survey

AIC – Akaike Information Criteria

AICc – Akaike Information Criteria (corrected)

ATA – Area-to-Area

ATP – Area-to-Point

GIS – Geographical Information Systems

GWR – Geographically Weighted Regression

GWPR – Geographically Weighted Poisson Regression

LISA – Local Indicators of Spatial Association

MAUP – Modifiable Areal Unit Problem

MSS – Mean Sum of Squares

RSS – Residual Sum of Squares

TABLE OF CONTENTS

INDEX OF TABLES

INDEX OF FIGURES

1. INTRODUCTION

1.1. Background

Geostatistics is a field traditionally dominated by the analysis of environmental datasets. By their nature, geostatistical methods are designed for spatially continuous attributes. Within the social sciences, data frequently represents a count instead of a continuous attribute, and is often only available at the administrative unit level. A typical approach to interpolating areal data is to use the centroids of the areas and perform point interpolation to create a continuous surface. This approach assumes that the data is located only at this one point, and not across the entire area. This assumption may only be valid when areas are regularly shaped and small, which is often not the case with social data.

Current visualisation methods for areal social data are typically based on Choropleth mapping. This method has a number of limitations. Rates can be unreliable in sparsely populated areas, and there may be visual bias when areas are of different shapes and sizes. There may also be a mismatch with spatial units for explanatory variables in regression modelling.

This thesis aims to contribute to existing work in the field of crime analysis by examining geostatistical methods for analysing areal data. Previous work on the geostatistical analysis of areal social data is largely theoretical, and this application will add a case study to the limited existing literature on this topic. The methods used, whilst employed on a crime dataset, could be applied to other similar areal social data featuring counts or rates.

The study explores how a new approach to the interpolation of areal data can be applied to analysing and visualising residential burglary data for the city of San Francisco, United States. Following on from this, the study aims to investigate the relationships between socio-economic variables and levels of crime. Both the original and interpolated data are used as inputs, in order to analyse how the interpolation method affects the results.

1.2. Objectives

In order to investigate approaches to analysing and visualising areal crime data, this study has been divided into the following objectives:

- Compare Area-to-Area and Area-to-Point kriging to Choropleth mapping and the traditional centroid method for interpolating residential burglary rates in the study area
- Locate spatial clusters of high or low residential burglary rates in the study area
- Explore the relationship between residential burglary and socio-economic variables in the study area using both non-spatial and spatial regression techniques

1.3. Assumptions

This study was motivated by the main hypothesis that commonly used visualisation methods for areal social data, such as Choropleth mapping and point kriging from geographical centroids, have a number of limitations. In addition, the following research assumptions are considered:

- The aggregation of data into areal units of different shapes and sizes can create a visual bias
- Area-to-Area and Area-to-Point kriging can be applied to analyse and visualise crime rate data
- It is possible to locate clusters of high or low residential burglary rates in the study area
- Socio-economic variables might influence residential burglary rates in the study area

1.4. General methodology

The methodology of this study comprises four main stages. The first stage is the literature review, which explores the key findings of the relevant literature. The second stage is exploring the data and devising an appropriate methodology in order to address the objectives of the study. The third phase is the analysis, in which the following methods are used: Choropleth mapping, point kriging from geographical centroids Area-to-Area and Area-to-Point kriging, Local Indicators of Spatial Association, and spatial and non-spatial regression. The final stage is evaluating the results and making conclusions with regard to the objectives of the study.

1.5. Dissertation organisation

This study is comprised of five chapters. The first chapter is the introduction, which includes a brief background to the study, along with the objectives, assumptions and the general methodology. The second chapter, the literature review, explores aspects of crime analysis relevant to the study; and the methods to be employed in this study to analyse and visualise areal data. The data and methods used in the study are described in the third chapter. In the fourth chapter the results of the study are presented and discussed. Finally, the fifth chapter details the conclusions of the study, outlines the limitations of this research and presents recommendations for further work.

2. LITERATURE REVIEW

The literature review explores the key findings of the relevant literature. It is comprised of two main sections. The first part describes aspects of crime analysis relevant to the study. Theories of crime and place, crime mapping methods and burglary are covered in order to provide a solid theoretical background.

The second section deals with the methods to be employed in this study to analyse and visualise areal data. Both geostatistical and non-geostatistical techniques are described, and their suitability for the study is explained. Relevant examples of applications of these methods are also mentioned.

2.1. Crime

A crime is defined as an act that breaches the criminal laws of an authority (such as a state or country). Crimes can be carried out against individuals, organisations, the state or involve the destruction of property. Within the United States, there are two classifications of crime: felonies and misdemeanours. A felony is defined as a serious crime (punishable by imprisonment of more than a year, or by the death penalty), whereas a misdemeanour is less serious.

Crime data, in conjunction with Geographical Information Systems (GIS), is commonly used to map and show visual patterns, look for clusters and explore relationships or causes. The relationship between crime and place or space as well as introductions to crime mapping and analysis are detailed in the sections below.

2.1.1. Crime and place

Crime and place are intertwined in a complex relationship. It is widely accepted by criminologists that crime occurs when there is an intersection of potential targets and offenders in space and time (Anselin et al., 2000). Analysing the relationship between crime and place can help to develop effective interventions to counter criminal behaviour in locations that experience high levels of crime.

Theories of the causes of crime and knowledge of effective crime reduction practices have informed concepts of crime and place (Anselin et al., 2000). Place may influence crime either by shaping the behaviour of people in an area, or by attracting people with criminal intentions to a location. In addition, spatial features such as public facilities can influence crime levels or types.

A number of criminological theories relate place and crime. One such theory is Routine Activities Theory. Within this theory, place is seen as facilitating crime in two ways. Firstly, the built environment can affect the capacities of crime suppressors. For instance, high-rise housing may decrease the monitoring by residents of public spaces at the street level. Secondly, the crime that occurs at a particular location is determined by the routine activities that occur there. Routine activities bring together potential offenders with opportunities to offend in particular locales. These locations may provide an abundance of opportunities for crime, or alternatively they may be locations of legal activities or facilities which are associated with an increased risk of crime, such as markets or other crowded places (Anselin et al., 2000).

A second concept that relates crime with place is the theory of "hotspots" (Sherman et al., 1989). A hotspot is an area with a high level of crime. Hotspots can be at a specific address, along a particular street, or at a neighbourhood or larger scale (Eck et al., 2005). The notion of crime hotspots hypothesises that certain land uses and social characteristics are related to high levels of crime. Within this theory, the built environment is also seen as influencing the level of crime, with signs of vandalism or disorder increasing the chance of more serious crimes. Crime hotspots may originally be concentrations of less serious crimes that later become hotspots of more serious crimes (Anselin et al., 2000). Knowledge of crime hotspots influences the behaviour of people, in the choices they make, from which areas to frequent or avoid, to how to interact with strangers (Eck et al., 2005). Knowledge of hotspots is frequently used by law enforcement authorities in order to allocate resources.

2.1.2. Crime mapping

Law enforcement officers and crime analysts were mapping crime well before the invention of GIS. Initially, mapping was carried out using pins and a paper map. Computerised crime mapping using GIS became widely available and affordable for crime agencies from the late 1980s onwards (La Vigne & Groff, 2001). GIS is now used to analyse incidents and to look for spatial or temporal trends in datasets which usually consist of large numbers of geographically referenced point records.

There are two competing viewpoints about the role of crime mapping within criminology. The first postulates that crime mapping can add to criminological theory by helping to make inferences about the possible processes underlying crime and crime patterns (Turton & Openshaw, 2001). The alternative stance states that knowledge of criminology theory is needed in order to interpret crime maps correctly (Pease, 2001). It has also been argued that crime mapping plays both roles, depending on the situation and user (Bowers & Hirschfield, 2001).

Many different types of crime map exist. The choice of which type to produce is guided by the type of data available (incident data, or aggregated data) and the end-purpose of the map. Common methods of crime mapping include point mapping, ellipse hotspot maps, Choropleth maps, interpolated or smoothed maps and isoline maps**.** All of the above methods have advantages and disadvantages. The benefits and limitations of perhaps the most common form of mapping for aggregated data, the Choropleth map, are discussed in detail later in this study.

Crime maps enable the visualisation and interpretation of data in an accessible way. However, researchers or organisations producing crime maps should be aware of the ability of maps to influence people in a way that may be more powerful and open to misinterpretation than the raw data (Bowers & Hirschfield, 2001). For example, areas incorrectly labelled as hotspots, whilst probably receiving more police resources, would be likely to suffer from falling property prices, increased insurance premiums, problems attracting highly-skilled employees and other secondary effects (Ratcliffe, 2002). Wallace (2009) suggests that the online crime mapping applications that are now commonly provided by police authorities portray crime as governable by reducing it to a series of symbols and administrative boundaries. The author also argues that such applications provide citizens with a means to judge local dangers and take on an element of responsibility for their own safety.

2.1.3. Crime analysis

Crime analysis is a practice that aims to identify patterns and trends in crime data. It can help to reconfigure police patrols, allocate public resources or decide where to locate police stations. Crime analysis and the resultant visual aids can also help to educate the public and promote community action (ESRI, 2008).

Early crime analysis methods frequently involved the aggregation of crime to areal units. Analysts compared variations in crime rates between different areas and looked for correlation between crime rates and social conditions (La Vigne & Groff, 2001). Modern crime analysis methods utilise GIS technology to identify risk factors and locations that may attract crime. GIS can also aid in the prediction of offender behaviour, based on previous crime trends. In addition, GIS may help identify suspects for series of crimes, from databases of previous offenders. In conjunction with secondary datasets, models of journeys to crime and causal relationships of crime can also be investigated.

In the United States, some crime information collected by public agencies is classified as public records, meaning that it is more freely available than in other countries (Ratcliffe, 2002). As with crime mapping, the choice of crime analysis method depends on the type of data available. For point crime incident data, a number of basic crime analysis methods include statistical tests to find patterns in the data. These include mean centre, standard deviation distance, standard deviation ellipse, and clustering tests (Eck et al., 2005). For aggregated areal data, the choice of traditional method is limited. Generally, Choropleth maps are created and analysed visually, or the rates are compared to local social and environmental conditions using geographical or non-geographical regression methods.

2.1.4. Burglary

Burglary is described as the act of entering a location (such as a dwelling or business) with the intent to commit larceny. Larceny is the obtaining of another party's property in an unlawful manner. According to the California Penal Code, Section 459 (2010):

"Every person who enters any house, room, apartment, tenement, shop, warehouse, store, mill, barn, stable, outhouse or other building, tent, vessel, floating home, railroad car, cargo container, trailer coach, house car, inhabited camper, vehicle, aircraft, or mine or any underground portion thereof, with intent to commit grand or petit larceny or any felony is guilty of burglary."

This study focuses on residential burglary. Residential burglaries are typically committed by groups (Mullins & Wright, 2003). Nee and Meenaghan (2006) state that many burglars employ searching strategies when looking for a residence to burgle, with crimes frequently not planned in advance.

A number of theories and studies that relate violent crime and place identify types of locations where offences are most likely to occur. According to Brantingham and Brantingham (2008), some places are "crime attractors", and others are "crime generators". The former are places where people work, shop or spend leisure time, and there is little or no effective policing. Crime generating areas are typically those with social problems and residential instability. Burglary is likely to occur in both types of locations, with different types of offenders and *modi operandi*.

Criminological studies have suggested many motivations for offenders to commit burglaries. Hearnden and Magill (2004) state that possible motivations include the need to fund drug use or buy alcohol. The unemployment rate of an area is also considered to influence the burglary rate, although the nature of the relationship has been debated (Deadman, 2003; Malczewski & Poetz, 2005). Other factors with unclear correlations with burglary rates include the percentage of the population without income and average household income (Malczewski & Poetz, 2005). Possible non-financial aspects related to higher burglary rates include low levels of education, and the percentages of lone parent households and multi-family dwellings (Malczewski & Poetz, 2005).

The motives above predominately relate to the characteristics or motivations of offenders. In addition, a number of factors relating to the residence have been identified. These include characteristics relating to the likely material benefits of a burglary, such as the value of a dwelling or a perception of relative wealth. However it has been found that offenders commit offenses based on the belief that there are valuable goods inside, instead of visible indications on the outside of the dwelling (Ham-Rowbottom et al., 1999; Hearnden & Magill, 2004; Malczewski & Poetz, 2005; Nee & Meenaghan, 2006). A non-financial aspect of property which may influence burglary rates is the proximity of neighbours. Ham-Rowbottom et al. (1999) note that dwellings which are overlooked by others (those in areas with high housing density) are less attractive to offenders.

Other factors influencing burglary rate include those related to distance from or access to properties. It has been found that burglaries often occur near the home address of an offender, for a number of reasons. These include the advantage of local knowledge and the impracticality of walking far whilst carrying heavy objects (Bernasco & Luykw, 2003; Hearnden & Magill, 2004). The principle of least effort suggests that criminals are more likely to commit offences near home or their place of work due to lower costs of travel in time and money. Bernasco and Luykw (2003) also propose that proximity to the central business district increases residential burglary rates. Breetzke (2012) postulates that physical geography may influence crime rates, and that mean altitude or slope may affect crime rates, although this hypothesis has yet to be thoroughly investigated.

Other influences on burglary rates that are non-spatial at the city-level relate to policing and the legal system. They include probability of conviction, the length of prison sentence, probability of imprisonment, and the number of police. All of which are likely to have a negative correlation with burglary rates (Deadman, 2003).

Spatial features such as roads, rivers and railway lines can act as delimiters of crime hotspots (Laukkanen et al., 2008; Van Patten et al., 2009). Other features and factors such as bars, off-licences, the level of traffic, land use, density of commercial activity are also associated with the level of crime in an area (Duffala, 1976).

Residential burglary is suitable for applied spatial research into crime because it usually occurs in a high enough volume to provide statistical significant results.

Burglaries also tend to cluster, which makes this crime type suitable for hotspot-type analyses.

2.2. Analysis methods for areal data

Crime data is commonly only available to those outside the police force in aggregated, areal units such as administrative areas. This is due to the need to uphold the privacy of both victims and offenders. However, when performing spatial crime analysis it is necessary to balance the requirements for both privacy and fine geographic resolution. Census tracts are seen to be an appropriate level for examining the relationship between neighbourhoods and crime. They have relatively homogeneous populations and a substantial level of data availability (Zhu et al., 2006).

The analysis of aggregated crime rate data poses a number of challenges. There are two main problems posed by rates calculated from low counts or populations. Firstly, variation in population means that the assumption of homogeneity of error variance is violated, as rates in areal units with a low population will have larger errors (this is known as the "Small Number Problem"). Secondly, error distributions cannot be assumed to be normal or symmetrical. As Osgood (2000) states, with low populations, rates of zero may be common and may potentially bias any regression results. High variability between rates may be the result of the small number problem. However, large differences between rates in areas with low populations may also be due to chance (Diehr, 1984). In areas with large populations even small differences in rates are likely to produce statistically significant results.

Aggregated data is usually displayed using Choropleth maps. This type of map, whilst being simple to produce and interpret, has a number of limitations. For example, in areas that experience few crimes or have a low population, there may be a large variation in numbers from one reporting period to another. This can lead to inaccurate or misleading maps (Williamson et al., 2001).

Crime rates are heteroscedastic, meaning that the variance at each location is a function of population. As population varies at each location, the standard errors of many types of statistic will not be constant over space. In addition, such areas are more likely to have rates that vary from the true rate (Cressie & Read, 1989).

The aggregation of data into areal units of different shapes and sizes can cause a visual bias (Goovaerts, 2006b). It is important to be aware of the effects of changing the values of categories when creating Choropleth maps, as the amount of crime can appear to be higher (or lower) than the actual situation (Williamson et al., 2001). Additionally, the different spatial supports for rates and explanatory variables may hinder correlation analyses such as regression methods (Goovaerts, 2006b).

The support of data refers to the size, shape and spatial orientation of the units of data. The problems associated with the changes in variance in the process of changing between supports are referred to as the "Change of Support Problem" (Gotway & Young, 2002). Geostatistical methods which may provide solutions to this problem include Block kriging, Point kriging, and cokriging. Non-geostatistical modelling approaches include scale-independent statistics, Multiscale Spatial Tree Models and Bayesian Hierarchical Models.

Mechanisms that play an important role in a process or phenomenon at one scale may not be relevant at another scale, and relationships may be obscured due to the choice of scale for analysis (Gotway & Young, 2002). The term "Modifiable Areal Unit Problem" (MAUP) was coined by Openshaw and Taylor (1979). MAUP consists of two interconnected issues. Firstly, the scale or aggregation effect considers different inferences obtained when data is regrouped into larger areal units. Secondly, the grouping or zoning effect concerns the variation in results obtained when the boundaries of areal units are placed differently, but still at the same scale (Gotway & Young, 2002; Openshaw & Taylor, 1979). The effects of MAUP are application and data specific, and are difficult or impossible to determine. Gotway and Young (2002) suggest that one way to overcome MAUP is to relate variation between aggregated areal units to the variation among the original units.

An additional challenge posed by aggregated data is that autocorrelation within an areal unit may be complex. Scales of variation may be different, or variation may by anisotropic or spatially heterogeneous (Haining et al., 2010). The process of aggregation leads to higher levels of spatial autocorrelation (Gotway & Young, 2002).

Despite the disadvantages of analysing aggregated areal data detailed below, studies at the aggregated areal unit scale have the benefit over the individual level in that such data is collected more routinely (Goovaerts, 2005).

2.2.1. Geostatistical methods for areal data

A number of geostatistical methods have been developed to try to address the aforementioned challenges of analysing areal count data. Goovaerts (2005) described three geostatistical methods to deal with the non-stationarity of variance caused by differing population sizes. Firstly, rates can be transformed, and subsequently a traditional geostatistical analysis can be carried out. This approach is limited as it is not possible to quantify the uncertainty associated with the transformed rates. Secondly, the population size can be incorporated into the semivariogram. However, such methods filter both variability caused by both the population-size effect and the underlying rate. A third approach is to develop new semivariogram and kriging algorithms which take into account the distributional nature of the count data. Binomial cokriging and Poisson kriging are examples of this.

2.2.1.1. Poisson Kriging

Poisson kriging is a variant of kriging in which count or rate data are interpreted as realisations of a random variable with a Poisson distribution. The rate (such as the crime rate or risk) at a particular location is estimated as a linear combination of the kernel rate and the rates observed in the neighbouring areas.

In the context of crime analysis, the Poisson distribution was originally derived from analysis of conviction rates in the 1820s in France (Osgood, 2000). The distribution provides the probability of observing a discrete number of events, given a mean count or rate of such events, occurring in a fixed time interval. It assumes that events occur independently of the time since the last event, and that the mean and variance of the distribution are equal. When the mean count is low, the distribution is skewed, but with higher means, the distribution approximates the Normal of Gaussian distribution. When populations are small there are only a limited number of probable rates. Higher populations produce smaller ranges of probable rates.

The first implementations of Poisson kriging assigned rates to the geographic centroids of areal units. This method assumes that all the inhabitants of an area live at this one point, and all crimes occur at this point too. This is only a suitable assumption if units are small compared to the interpolation grid (Goovaerts, 2006b).

The Poisson kriging method introduced by Goovaerts (2005) enables the modelling of spatial correlation of rate data. Spatial dependence is taken into account when estimating the underlying risk behind the rates, and the associated uncertainty. In this process, population size is also factored in.

One disadvantage of Poisson kriging is that the uncertainty associated with the parameters of the correlation function is not taken into account. In Full Bayesian Modelling, they are included in the analysis. This will probably mean smaller prediction variances for Poisson kriging (Goovaerts, 2005). Area-to-Area and Area-to-Point Poisson kriging aim to deal with the aforementioned limitations of Poisson kriging. Such methods are discussed below.

2.2.1.2. Area-to-Area and Area-to-Point Poisson Kriging

An extension or improvement to Poisson kriging for rates in areas is Area-to-Area (ATA) Poisson kriging. This kriging method can incorporate different spatial supports for the data and prediction units. Geostatistics has long been used to obtain block average predictors from point data, although these are generally regularly shaped and sized blocks.

Area-to-Area Poisson Kriging calculates the approximate covariance between two areas. This is done by calculating the average of the point-support covariance between any two points discretising the two areas. However, the point-support variogram cannot be obtained directly from the data. The first step in obtaining the point-support semivariogram is to model the semivariogram of the areal data. This is then deconvoluted in order to obtain the point-support variogram (Goovaerts, 2006b). Deconvolution is described in more detail below.

The term Area-to-Point (ATP) kriging was coined by Kyriakidis (2004). ATP kriging is a specific form of ATA kriging, in which the prediction support is small enough to be treated as a point. When ATP kriging is carried out at all nodes of a grid or raster map, then a continuous surface can be produced.

Both ATA and ATP Poisson kriging, as introduced by Goovaerts (2006b), take into account the shape and size of areal units, in addition to their different population sizes. ATP kriging produces coherent predictions. This means that the sums of disaggregated estimates are non-negative and equal to the original aggregated data.

ATA Poisson kriging is most commonly used for the mapping of diseases. For instance, Goovaerts (2006b) uses ATA and ATP Poisson kriging to map real and simulated cancer mortality data. Goovaerts found that ATP kriging produces more accurate predictions and confidence interval than point kriging from the centroids of areal data. The author also found that as administrative units become more heterogeneous, the benefits of using ATP instead of point kriging from areal centroids increase.

Kerry et al. (2010) used ATA Poisson kriging to study crime data in large administrative units. ATA kriging was used to filter the noise in rates caused by the small number problem. The authors used ATP Poisson kriging to create continuous crime risk maps.

Crucial to the ATA and ATP kriging processes is the point-support covariance of crime risk, or the point-support semivariogram. The method for obtaining the semivariogram of the areal data is known as deconvolution. Deconvolution is frequently carried out in applications of geostatistics in the field of mining. In these applications, the units are typically the same shape and size, meaning that deconvolution is less complex.

Goovaerts (2008a) describes an adaptation of such methods in order to make them suitable for irregular geographical units. To begin with, an initial point-support model is chosen, which is then regularised using a regularisation expression. This is composed of two terms. The first of which is the semivariogram of values within a unit, which varies as a function of a separation vector. This is because the size of geographical units varies as a function of the distance between units. The second term is related to the semivariogram value between two units.

Following this, the theoretically regularised model is compared to the data-based model. To optimise the solution, the relative difference between the two semivariogram curves is used. Following an iterative procedure, new models are considered an improvement if there is less deviation between the theoretically regularised model and the areal data model.

This approach to deconvolution relies on the assumption that the average distance between units is representative. This may not be the case when there are extremely irregularly shaped units (Yoo et al., 2010).

2.2.2. Other methods for areal data

Statistical methods for regional data include spatial proximity measures, spatial smoothing methods (such as locally weighted averages or Bayesian smoothing or hierarchical modelling), cluster detection methods (such as Moran's I, Tango's index and Scan statistics) and multivariate analysis (Krivoruchko et al., 2003). Non-parametric spatial smoothing methods include spatial filtering and the head-banging algorithm, which are variations of a moving window kernel-based smoother (Johnson, 2004). The head-banging algorithm takes spatial geometry and the values of surrounding observations into account. However, anisotropy is not accounted for, and the range of spatial correlation and the uncertainty of smoothed rates cannot be quantified (Goovaerts, 2005). Three of these methods suitable for analysing areal crime data are discussed in more detail below.

2.2.2.1. Local Indicators of Spatial Association

Local Indicators of Spatial Association (LISA) were originally proposed by Anselin (1995). These indicators are designed for the decomposition of global indicators to find the contribution of each observation. One of the most commonly used of these indicators is a local version of Moran's *I*, a measure of spatial autocorrelation. It can be used to identify local spatial clusters and find spatial outliers in global measures of spatial association.

Many studies of crime have used LISA in order to identify clusters (or hotspots) of crime. Almeida et al. (2005) use LISA to explore the spatial patterns of crime in part of Brazil. Statistically significant spatial clusters of crime are located using a local version of Moran's *I*.

Kerry et al. (2010) built on the results of ATA and ATP Poisson kriging of aggregated crime data by using them as an input for LISA. Rates were simulated using *p*-field simulation. Local Moran's *I* was then calculated for each of the simulated rate maps. Significance was tested using randomisation procedures. This method was carried out using the original, ATA and ATP kriged rates. For each of these types of rates, a summary was produced. This showed the category under which each unit was most frequently classified in the simulated rate maps. For the data used in the study, the ATA kriged rates show fewer significant clusters of high or low crime rates. ATP kriged rates show more clusters, as would be expected due to the difference in scale. The authors suggest that performing ATP kriging before LISA cluster analysis may be beneficial for accurately locating where clusters may be. This procedure aims to allow for the rejection of the null hypothesis that any variation is spatially random.

A similar procedure was used to analyse health data in the study by Goovaerts and Jacquez (2005). The authors use sequential Gaussian simulation to generate multiple realisations of the spatial distribution of mortality rates under a variety of conditions. These are used to produce 'neutral models', reflecting a plausible scenario of background variation, as some level of spatial dependency is to be expected. These simulated neutral models are analysed using spatio-temporal variants of Moran's *I* statistic.

Goovaerts (2006a) used Poisson kriging, in combination with non-conditional Gaussian simulation to generate multiple realisations of the spatial distribution of the variable concerned. Local Moran's I was then used to detect local clusters in the data.

2.2.2.2. Areal regression

Regression methods for areal crime data include both geographical and non-geographical methods. Linear regression models are not adequate to describe the relationship between count or rate data and explanatory variables. This is because rates, especially those based on low counts or populations, cannot be approximated as continuous variables and therefore linear regression models are unsuitable (Gotway & Wolfinger, 2003).

Generalised linear models are frequently used to overcome this problem. Poisson regression models are a form of generalised linear model, and aim to overcome challenges posed by the use of counts or rates. Such models incorporate population into the model equation in order to analyse rates. This aims to deal with the aforementioned problem of the heterogeneity of error variance, as it recognises the higher precision of rates in areas with higher populations (Osgood, 2000).

Analyses of crime rate data commonly experience problems of overdispersion (Osgood, 2000). Overdispersion is said to exist when the variance is higher than the mean, thus violating a key attribute of the Poisson distribution. Methods aiming to overcome overdispersion include the quasi-likelihood approach, negative binomial regression and methods with a case-specific residual term, similar to an error term.

Negative binomial regression is commonly used to address overdispersion in criminological studies (Berk & MacDonald, 2008). The negative binomial distribution, similarly to the Poisson distribution, is a discrete probability distribution. The negative binomial regression model has an additional term which reflects unexplained variation in the underlying mean event counts or rates.

However, despite its aim, the negative binomial model can still have problems with overdispersion (Berk & MacDonald, 2008; Law & Haining, 2004).

2.2.2.3. Geographically Weighted Regression

Geographically Weighted Regression (GWR) is a form of regression that permits parameter estimates to vary locally. It was developed by Fotheringham et al (2002). Weights are used in the regression so that the nearest observations have the highest weighting. It has been suggested by Fotheringham (2002), that the results of GWR might be more robust to scale issues than global results. Evidence of spatial autocorrelation in the residuals of a global regression model could provide justification for the use of this method.

Goovaerts (2008b) utilised GWR to analyse the results of ATA and ATP Poisson kriging of mortality risk data. GWR is used to show how well socio-economic variables explain variation in the kriged dependent variable.

However, there has been criticism of GWR as a modelling tool. Tiefelsdorf and Wheeler (2005) argue that any multicollinearity within the data may be increased by calculating local GWR coefficients. The results of their study indicate that local regression coefficients can be collinear even if the underlying variables in the data are uncorrelated.

3. DATA AND METHODS

The following section describes the study area, the data, and the theoretical background for all the methods to be used in the analysis. It is divided into five sections. The first briefly introduces the study area and the data used in this study. The following four sections each summarise a different method and state the relevant formulas.

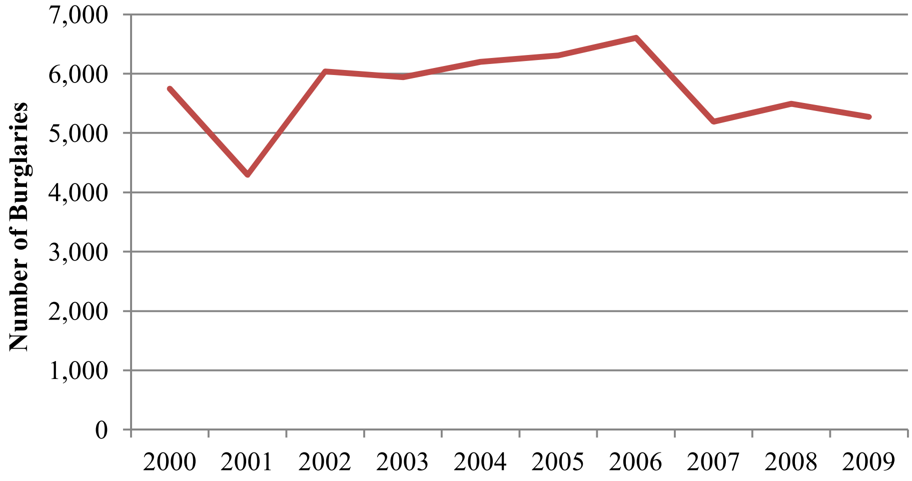
3.1. Study area and data

This section provides a brief introduction to the study area for this research project; San Francisco, a city and county located in California, United States. The city is part of a larger populated area which also includes the cities of Oakland, and San Jose, along with smaller settlements. In 2010 San Francisco had a population of 805,235 (U.S. Census Bureau, 2010a). The area first experienced growth with the gold rush in the late 1840s. In the following years, large numbers of both national and international migrants settled in the city. San Francisco has an ethnically diverse population, with 33% Asian, 15% Hispanic, and 6% Black or African American residents (U.S. Census Bureau, 2010b).

Since the 1990s, San Francisco has expanded to become a regional centre for technology and finance. The unemployment rate in November 2011 was 8.1%, below the rate of 9% for the United States (State of California, 2011). Household incomes in the city are high compared to the national average; while levels of poverty are lower than average.

3.1.1. Crime in San Francisco

Whilst crime rates in San Francisco have steadily decreased since 2000, the burglary rate has stayed fairly stable. Figure 1 shows the number of burglaries in San Francisco County in each calendar year from 2000 to 2009.



Source: California Department of Justice

Figure 1: Burglary trend in San Francisco County, 2000-2009

The profiles and ages of offenders in San Francisco are varied. According to Males (2009), residents of San Francisco aged between 50 and 59 years commit more crime than those under 18 years old, and those aged between 40 and 49 commit three times as many crimes. Males also claims that murder rates are high in areas with poverty levels of over 20 percent.

3.1.2. Data

The dataset used in this study is reported burglaries in San Francisco in 2010. The dataset was obtained from the San Francisco Police Department website. It comes without metadata, and therefore there are no assurances of data quality or consistency of data collection.

The focus of the study is residential burglary; therefore all non-residential burglaries were removed from the dataset. The original point data was aggregated to the census tracts of San Francisco City (as of 2010), excluding islands falling within the boundaries of the city. This leaves 194 census tracts of irregular shapes and sizes in the study area.

The aggregated counts were converted to rates of burglaries per 1000 housing units using housing unit counts from the 2010 Census. A housing unit is a house, mobile home or trailer, apartment, group of rooms, or single room that is occupied (or intended to be occupied) as a separate living quarters (U.S Census Bureau, 2000). The results of this aggregation are shown in Figure 2. It was felt that the number of housing units, or residences, would more accurately reflect the risk of burglary than the more traditional method of calculating rates using population data, as the number of people per housing unit is likely to vary greatly across the study area.

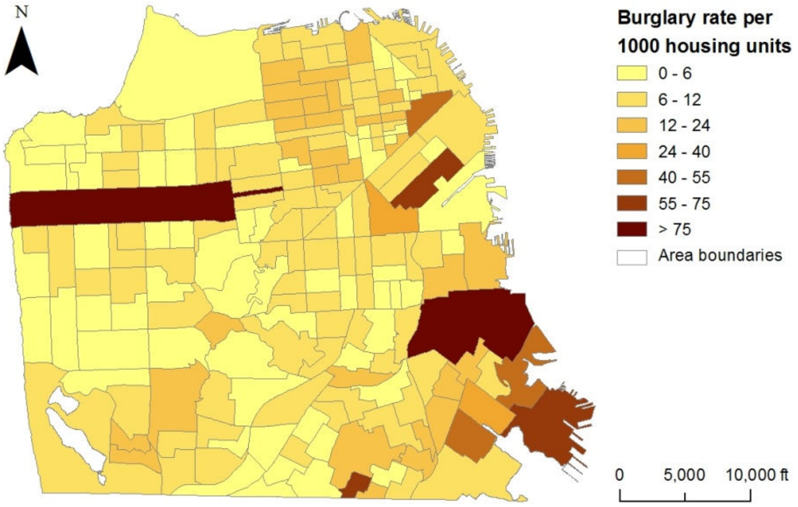


Figure 2: Burglary Rate per 1000 housing units

A summary of the residential burglary counts and the calculated burglary rates in the census tracts is presented in Table 1 below.

Table 1: Summary of burglary counts and rates for the census tracts

	Mean	Variance	Min	Max
Burglary counts	17.995	188.088	0	99
Burglary rates	11.910	451.262	0	206.250

Also used in the analysis were housing unit counts for census blocks (an administrative unit smaller than census tracts), also from the 2010 Census.

The datasets of the socio-economic variables (Table 2) used in the areal regression were obtained from the 2010 American Community Survey (ACS), which is conducted by the US Census Bureau. The drug rate dataset was constructed using geocoded drug incident data from the San Francisco Police Department and population data from the 2010 Census. The housing density measure of housing units per hectare was calculated in ArcGIS using 2010 Census data and census tract boundaries.

Table 2: List of socio-economic variables for regression

Variable	Original sources	Year
Education (percentage of population 25 years and over with less than a high school education)	ACS	2010
Drug incident rate per 1000 residents	San Francisco Police Department, 2010 Census	2010
Household income (median household income in the past 12 months, in thousands of 2010 inflation-adjusted dollars)	ACS	2010
Housing density (housing units per hectare)	2010 Census	2010
House value (percentage of owner occupied homes with value over 500,000 dollars)	ACS	2010

3.2. Poisson Kriging

The burglary count $d(\mathrm{u}_\alpha)$ was interpreted as a realisation of a random variable $\mathrm{D}(\mathrm{u}_\alpha)$ that has a Poisson distribution with the one parameter (expected number of burglaries). This parameter is the product of the number of housing units $n(\mathrm{u}_\alpha)$ and the local risk of burglary (or noise-filtered burglary rate) $R(\mathrm{u}_\alpha)$. The noise-filtered burglary rate was estimated using Poisson kriging, which aims to filter the noise associated with the rates of areas with low populations. Poisson kriging using the conventional method of collapsing areas to their centroids is described in the following paragraphs.

The burglary risk ($r(\mathrm{u}_\alpha)$), at location (u_α) is estimated by

$$\hat{r}(\mathrm{u}_\alpha) = \sum_{i=1}^{K} \lambda_i(\mathrm{u}_\alpha) z(\mathrm{u}_i) \qquad \textbf{Equation 1}$$

where $\lambda_i(\mathrm{u}_\alpha)$ is the weight assigned to the rate $z(\mathrm{u}_i)$ when estimating the risk at (u_α). The associated kriging (prediction) variance is computed using the following formula:

$$\sigma^2(\mathrm{u}_\alpha) = C_R(0) - \sum_{i=1}^{K} \lambda_i(\mathrm{u}_\alpha) C_R(\mathrm{u}_i - \mathrm{u}_\alpha) - \mu(\mathrm{u}_\alpha) \qquad \textbf{Equation 2}$$

The *K* weights and Lagrange parameter are calculated by solving the following system of (*K*+1) linear equations:

$$\begin{aligned}
&\sum_{j=1}^{K} \lambda_j(\mathrm{u}_\alpha) \left[C_R\left(\mathrm{u}_i - \mathrm{u}_j\right) + \delta_{ij} \frac{m^*}{n(\mathrm{u}_i)} \right] + \mu(\mathrm{u}_\alpha) = C_R(\mathrm{u}_i - \mathrm{u}_\alpha) \quad i = 1, \dots, K \\
&\sum_{j=1}^{K} \lambda_j(\mathrm{u}_\alpha) = 1
\end{aligned} \qquad \textbf{Equation 3}$$

where $\delta_{ij} = 1$ if $\mathrm{u}_i = \mathrm{u}_j$ and 0 otherwise. $n(\mathrm{u}_i)$ is the number of housing units at (u_i), and m^* is the housing-unit-weighted mean of the *N* rates. $m^*/n(\mathrm{u}_i)$ is an error variance term, meaning that areas with fewer housing units have a lower weight.

The semivariogram of burglary risk, required by equations 2 and 3, is estimated as:

$$\hat{\gamma}_R(\mathrm{h}) = \frac{1}{2\sum_{\alpha=1}^{N(\mathrm{h})} \frac{n(\mathrm{u}_\alpha) n(\mathrm{u}_\alpha + \mathrm{h})}{n(\mathrm{u}_\alpha) + n(\mathrm{u}_\alpha + \mathrm{h})}} \sum_{\alpha=1}^{N(\mathrm{h})} \left\{ \frac{n(\mathrm{u}_\alpha) n(\mathrm{u}_\alpha + \mathrm{h})}{n(\mathrm{u}_\alpha) + n(\mathrm{u}_\alpha + \mathrm{h})} [z(\mathrm{u}_\alpha) - z(\mathrm{u}_\alpha + \mathrm{h})]^2 - m^* \right\} \qquad \textbf{Equation 4}$$

where *N*(h) is the number of data pairs separated by the vector h. The pairs $[z(\mathrm{u}_\alpha) - z(\mathrm{u}_\alpha + \mathrm{h})]$ are weighted by number of housing units in order to make their variance consistent.

3.3. ATA and ATP Poisson Kriging

For ATA and ATP Poisson kriging, burglary risk at a location is estimated by (Goovaerts, 2008b):

$$\hat{r}(X) = \sum_{i=1}^{K} \lambda_i(X) z(v_i) \qquad \textbf{Equation 5}$$

where *X* represents either an area (v_α) in the case of ATA kriging or a point $(u_s)$for ATP kriging.

Kriging variance is computed as follows:

$$\sigma^2(X) = \bar{C}_R(X,X) - \sum_{i=1}^{K} \lambda_i(X) \bar{C}_R(v_i, X) - \mu(X) \qquad \textbf{Equation 6}$$

The kriging weights and Lagrange parameter are computed by solving the following system of equations:

$$\begin{aligned} &\sum_{j=1}^{K} \lambda_j(X) \left[\bar{C}_R\big(v_i, v_j\big) + \delta_{ij} \frac{m^*}{n(v_i)} \right] + \mu(\mathrm{u}_\alpha) = \bar{C}_R(v_i, X) \quad i = 1, \dots, K \\ &\sum_{j=1}^{K} \lambda_j(X) = 1 \end{aligned} \qquad \textbf{Equation 7}$$

where δ_{ij}=1 if i=j and 0 otherwise. As with Poisson kriging from centroids as described above, $m^*/n(\mathrm{u}_i)$ is an error variance term, and areas with fewer housing units have a lower weight.

The main difference between ATA and ATP kriging and the centroid method is that the point-to-point covariances are replaced by area-to-area covariances. For ATA Poisson kriging, these are estimated as the average of the point support covariance *C*(h) between any two locations discretising the areas (v_i) and $\big(v_j\big)$:

$$\bar{C}_R(v_i, v_j) = \frac{1}{\sum_{s=1}^{P_i} \sum_{s'=1}^{P_j} w_{ss'}} \sum_{s=1}^{P_i} \sum_{s'=1}^{P_j} w_{ss'} \, C(u_s, u_{s'}) \qquad \textbf{Equation 8}$$

where P_i and P_j are the number of points used to discretise the two areas (v_i) and (v_j). The weights $(w_{ss'})$ are the product of the number of housing units within the cells focused on the discretising points (u_s) and $(u_{s'})$. See section 3.3.1 for a brief description of the construction of the housing unit map.

For ATP Poisson Kriging, area-to-point covariances are approximated by:

$$\bar{C}_R(v_i, u_s) = \frac{1}{\sum_{s'=1}^{P_i} w_{s's}} \sum_{s'=1}^{P_i} w_{s's} C(u_{s'}, u_s) \qquad \textbf{Equation 9}$$

One of the most important properties of ATP Poisson kriging is that estimates are coherent. The housing-unit-weighted average of the rates estimated at the discretisation points is equal to the ATA estimate.

For ATA and ATP Poisson kriging, knowledge of the point support covariance of the risk C(h) or the semivariogram of burglary risk is necessary. This cannot be estimated from the observed rates (as they are not available). Deconvolution is used to derive the point support semivariogram from the experimental semivariogram of areal data. When the areas are irregularly shaped and sized, Goovaerts (2008a) suggested an iterative approach, starting with the derivation of an initial deconvoluted model $\gamma^{(0)}(\mathrm{h})$. The initial model is regularised:

$$\gamma_{regul}(\mathrm{h}) = \bar{\gamma}^{(0)}(v, v_h) - \bar{\gamma}_h^{(0)}(v, v) \qquad \textbf{Equation 10}$$

where $\bar{\gamma}^{(0)}(v, v_h)$ is the area-to-area semivariogram value for two areas separated by distance *h*. This is approximated using equation 8, with $\gamma^{(0)}$ in the place of *C*(h). $\bar{\gamma}_h^{(0)}(v, v)$ is the within-area semivariogram value, which varies as a function of distance because smaller areas are paired at shorter distances. Variations between housing-unit density are accounted for by estimating the distance between two areas as a housing-unit-weighted average between the discretising locations.

The theoretically regularised model from equation 10, γ_{regul}, is compared to the model fitted to the experimental values $\gamma_R(\mathrm{h})$, and the relative distance between the two curves is minimised iteratively by rescaling the initial point-support model $\gamma^{(0)}(\mathrm{h})$.

ATA and ATP Poisson kriging was carried out using the SpaceStat software. The housing unit map described in section 3.3.1 was used as an input for ATA and ATP kriging and as the grid of cells for the ATP output. The map was also used to create the housing-unit-weighted centroids (see Figure A- 2). Thirty random discretisation points were used to construct the variograms. The spatial weighting used to produce the kriging estimates was one ring of queen neighbours, standardised by neighbour count.

3.3.1. Housing unit map

A map of the approximate number of housing units per grid cell was created as an input for ATA and ATP kriging. Housing unit density was assumed to be constant within each census block for the purposes of this analysis. Areal weighting was used to approximate the number of housing units in each unit of 300 feet by 300 feet. The resultant map can be seen in Figure A- 1.

3.4. Local Indicators of Spatial Association

The Local Indicator of Spatial Association used in this study is the Local Moran's *I* statistic. It is calculated using the following formula (Anselin, 1995):

$$I\,(\mathrm{v}_i) = \left[\frac{z(\mathrm{v}_i) - m}{s}\right] \times \sum_{j=1, j\neq i}^{n} w_{i,j} \times \left[\frac{z(\mathrm{v}_j) - m}{s}\right] \qquad \textbf{Equation 11}$$

where $z(\mathrm{v}_i)$ is the attribute, *m* and *s* are the mean and standard deviation of the set of areas and $w_{i,j}$ represents the spatial weight between features *i* and *j*.

Positive values of *I* indicate that neighbouring features are similarly high or low and that there is a cluster. Negative values indicate dissimilar neighbouring values, therefore the feature is an outlier. P-values distinguish statistically significant values of Moran's *I*. The type of cluster is then determined by comparing the local mean of a target feature's neighbours to the global mean. Features with a local mean higher than the global mean are classified as High-High clusters, or hotspots. Features with a local mean smaller than the global mean are deemed to be Low-Low clusters. The type of outlier is established by comparing the value of the target feature to the local mean. Features with values that are higher than the local mean are classified as High-Low outliers. Features with a value lower than the local mean are classified as Low-High outliers.

This Local Moran's *I* analysis was carried out in ArcMap 10 with the original data, the ATA and ATP results and the centroid method of Poisson kriging. Local Moran's *I* analysis requires the conceptualisation of the spatial relationships between features. The inverse distance squared conceptualisation was used for all data sets in this study. This method means that all features influence all other features, but only a target feature's closest neighbours will exercise substantial influence in the calculations for that feature. Distances are calculated from the polygon centroids.

3.5. Areal regression

This study investigates residential burglary rates using variables relating to the characteristics of the neighbourhoods of the victims, not necessarily the neighbourhoods of the offenders (although this may be the same area). As stated in Section 2.1.4, burglaries often occur near the home of an offender (Bernasco & Luykw, 2003; Hearnden & Magill, 2004). The same section also describes how characteristics relating to location, offenders and the properties may play roles in determining burglary rates. These aspects should be factored in when interpreting the results of the Poisson regression.

The basic model for Poisson regression is:

$$\ln(\lambda_i) = \sum_{k=0}^{K} \beta_k x_{ik}$$ **Equation 12**

$$P(Y_i = y_i) = \frac{e^{-\lambda_i}\lambda^{y_i}}{y_i!}$$ **Equation 13**

Equation 12 is the regression equation. The natural logarithm of the expected number of events for case *i,* relates to the sum of the products of each explanatory or dependent variable x_{ik}, multiplied by a regression coefficient, β_k. β_0 is a constant that is multiplied by 1 for every case. Equation 13 states that the probability of y_i (the observed outcome) follows the Poisson distribution. The expected distribution of counts and residuals depends on one parameter, the fitted mean count, λ_i. For rates, equation 12 is altered to (Osgood, 2000):

$$\ln\left(\frac{\lambda_i}{n_i}\right) = \sum_{k=0}^{K} \beta_k x_{ik}$$

$$\ln(\lambda_i) = \ln(n_i) + \sum_{k=0}^{K} \beta_k x_{ik}$$ **Equation 14**

where n_i is the population size, or number of housing units. The term $\ln(n_i)$ has a fixed coefficient of 1.

If data is overdispersed (residual variance exceeds λ_i), negative binomial regression can be used. It combines the Poisson distribution and a gamma distribution of unexplained variation in the underlying mean. Equation 13 is replaced by:

$$P(Y_i = y_i) = \frac{\Gamma(y_i + \phi)}{y_i!\,\Gamma(\phi)} \frac{\phi^{\phi}{\lambda_i}^{y_i}}{(\phi + \lambda_i)^{\phi - y_i}}$$ **Equation 15**

where Γ is the gamma function, and ϕ is the reciprocal of the residual variance of underlying mean counts. The gamma function is a continuous version of the factorial function.

Overdispersion in the residuals of Poisson regression was tested for using a test score developed by Dean (1992). Pseudo R-squared values were calculated as an indication of the goodness-of-the models. Cragg and Uhler's pseudo r-squared (Cragg & Uhler, 1970) is calculated using likelihood ratios, and ranges from 0 to 1.

Regression was carried out in R using the original data and ATA Poisson kriged data. This was done in order to investigate the effect of using interpolated data to form models. Goovaerts (2006b) suggested that using ATA kriged data may help to alleviate the problems caused by performing regression analysis at scales that may misrepresent the relationship between response and explanatory variables.

3.5.1. Geographically Weighted Regression

For Geographically Weighted Poisson Regression (GWPR) the equation is similar to equation 12, but the parameter estimates β_i^* are specific to each location *i*. The regression model is fitted using iteratively reweighted least squares.

Global Moran's *I* was calculated for the residuals of Poisson regression to check for spatial autocorrelation and determine if GWPR may be suitable. Software developed by the authors of Fotheringham et. al (2002) was used to carry out GWPR using housing-unit-weighted centroids as both the data and prediction points. Adaptive bandwidths were selected using an iterative process which minimises the Corrected Akaike Information Criterion (AICc), a measure of the relative goodness-of-fit of a model. Models with a smaller AIC or AICc score are considered to be better models. For more information on the AICc, see Sugiura (1978).

4. RESULTS AND DISCUSSION

The following section presents and discusses the findings of this study. It is divided into five parts, each detailing the findings of a specific technique of analysis. Additional figures can be found in Appendix A.

4.1. Poisson Kriging

The kriging estimates of Poisson kriging using the geographical centroids are presented in Figure 3. Concentric circular patterns around the centroids are clearly visible in a number of locations. These include high values in the Golden Gate Park to the west of the study area, and low values in the Presidio in the north. The area around Islais Creek (to the south-east of the study area) is where the highest values are found. Figure 4 highlights the names and locations of the neighbourhoods mentioned in this study.

The problems caused by using centroids are potentially most obvious in the Golden Gate Park, due to the elongated shape of the polygon and its large relative size. The influence of the high rate in the park extends to the north and south outside of the polygon, and does not cover much of the east-west extent. There are only 37 housing units located within the park, and these could be located at any position within the polygon. It is unlikely that this map provides a realistic depiction of reality in the park.

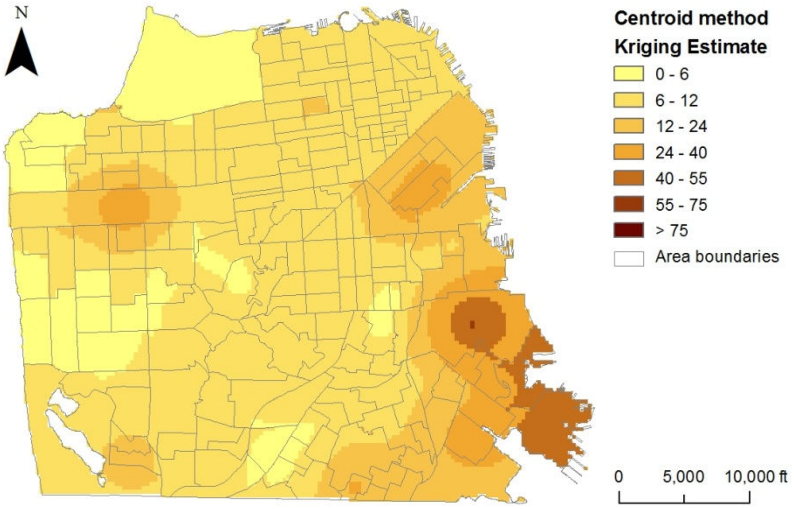


Figure 3: Burglary Rate Centroid Method Kriging Estimate

The limitations of the centroid method are also evident in the kriging variance map presented in Figure 5. In the Golden Gate Park, kriging variance is relatively low in the centre, and high in the east and west. Similarly, kriging variance is high in the Presidio and the area around Islais Creek.

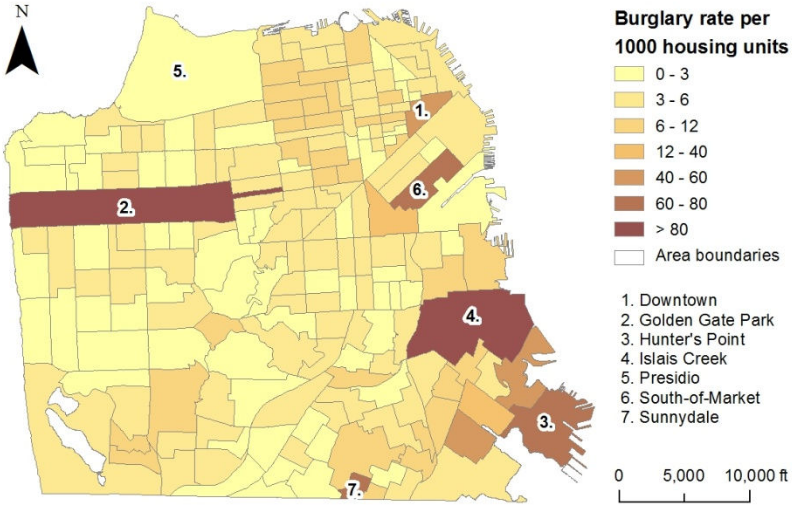


Figure 4: Selected city neighbourhoods

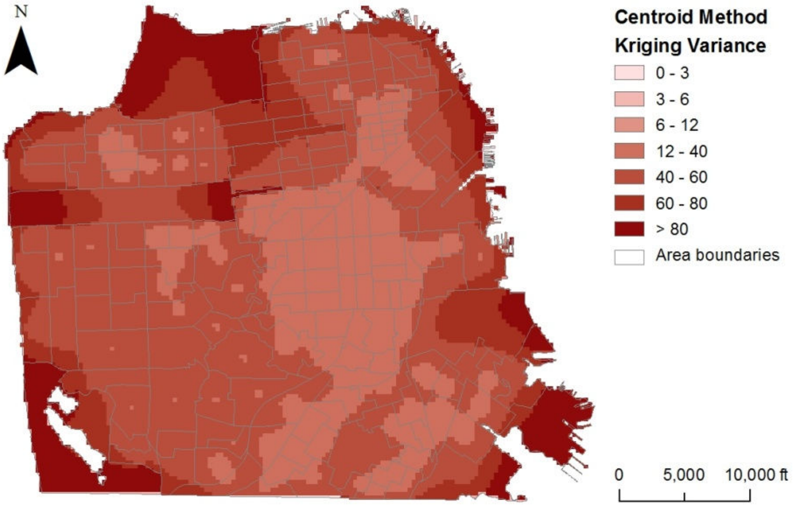


Figure 5: Burglary Rate Centroid Method Kriging Variance

The spherical variogram model (Figure 6) has a Mean Sum-of-Squares (MSS) error of 0.050. This is the error between the experimental variogram and the variogram model.

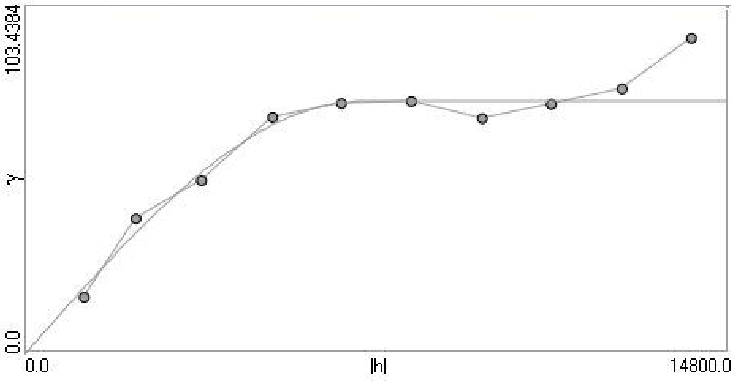


Figure 6: Centroid Method Burglary Rate Variogram

4.2. ATA and ATP Kriging

Following the creation of the housing unit dataset (see Figure A- 1), a housing-unit-weighted centroid dataset (Figure A- 2) was produced. These two datasets, along with the residential burglary dataset (Figure 2) were used as the inputs for ATA and ATP Poisson kriging.

For ATA and ATP kriging the variogram and deconvolution models are the same. The fitted exponential variogram model (Figure 7) has an MSS error of 0.012.

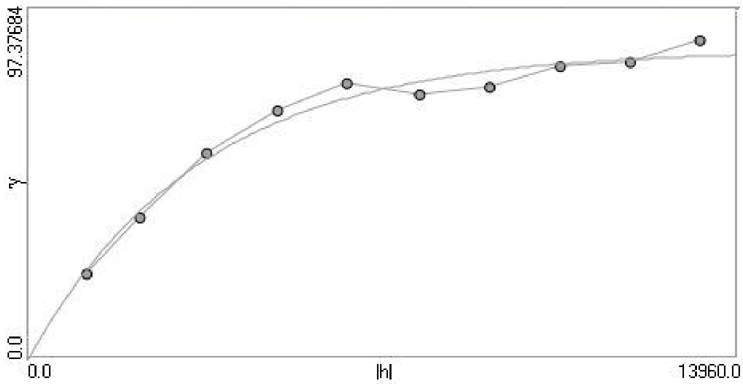


Figure 7: Burglary Rate Variogram

The exponential deconvolution model presented in Figure 8 has an MSS error of 0.024.

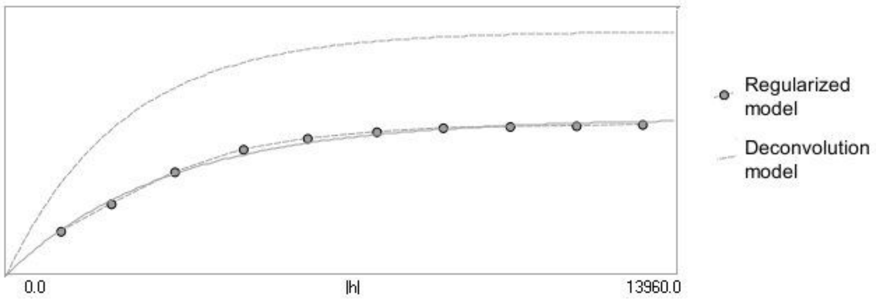


Figure 8: Burglary Rate Deconvolution model

The kriging estimate map for ATA Poisson kriging (Figure 9) shows a lower range of burglary rates than the original data. This is also visible in the summary table of all kriging results,

Table 3. The high original rates in the Golden Gate Park (189.18) and the area near Islais Creek (206.25) are reduced to 15.41 and 66.81 respectively. The decrease in rate for the Golden Gate Park is noticeable. This is likely due to the combination of a low number of housing units in the area and large area size. However, the area around Islais Creek still has the highest rate (although it is considerably lower).

Other areas with high rates include the areas south of Islais Creek (Hunter's Point) and an area near the South-of-Market neighbourhood, in the east of the study area.

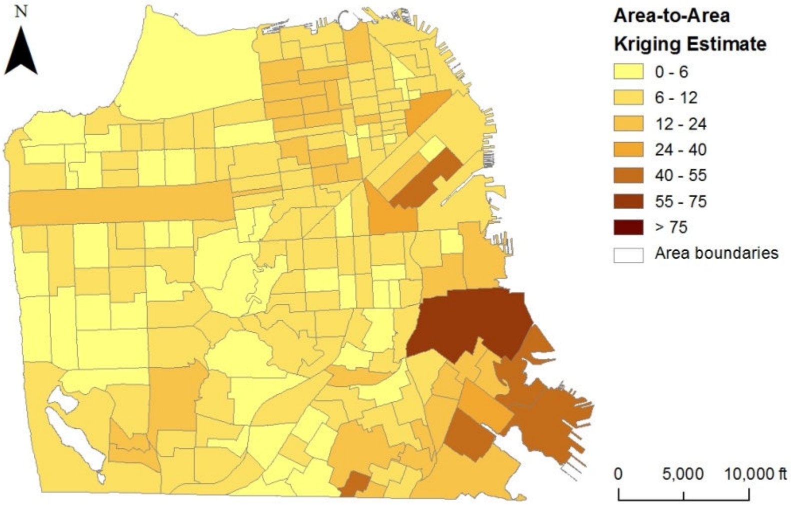


Figure 9: Burglary Rate Area-to-Area Kriging Estimate

As mentioned in previous paragraphs, kriging estimates in the areas with the highest burglary rates in the original dataset have decreased. This is clearly visible in the ATA kriging residual map shown in Figure 10. In addition, the map shows that in the vicinity of some polygons with high kriging residuals, kriging residuals are low. This means that kriging estimates in these regions have been smoothed. Evidence of this is mostly in the south-east of the study area.

Figure 11 presents the kriging variance map. The kriging variance is highest in the polygons with high or low kriging estimates. It appears that there is not a strong pattern linking polygon size with kriging variance, although some of the larger polygons have relatively large variances.

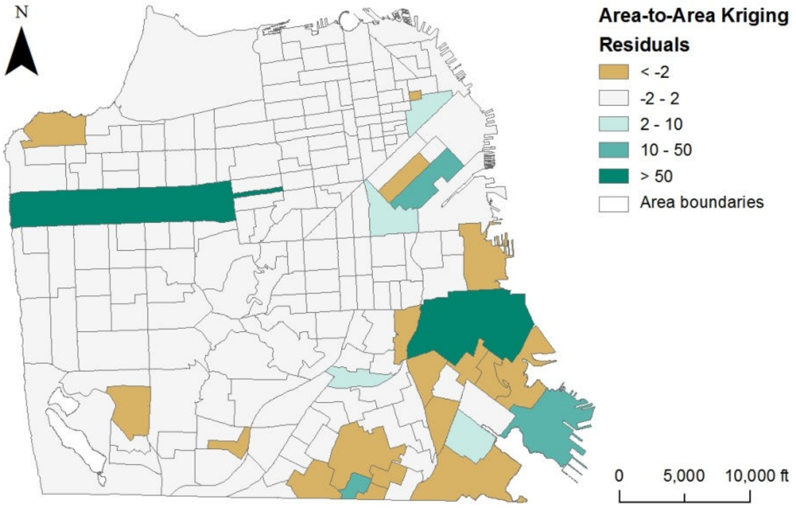


Figure 10: Burglary Rate Area-to-Area Kriging Residuals

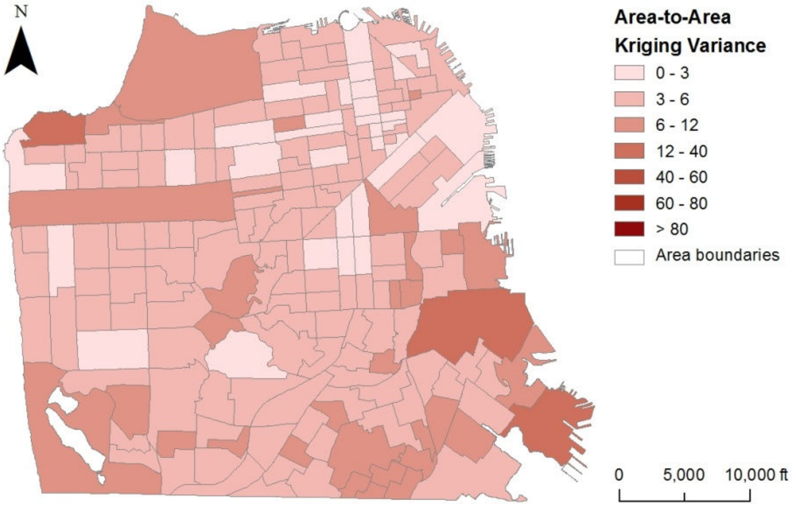


Figure 11: Burglary Rate Area-to-Area Kriging Variance

The correlation between the ATA Poisson kriged and original burglary rates is shown in Figure 12. Contrary to what would be expected, the correlation between the two datasets is low, with an R-squared value of 0.5109. However, removing the two highest values increases the R-squared value to 0.9631.

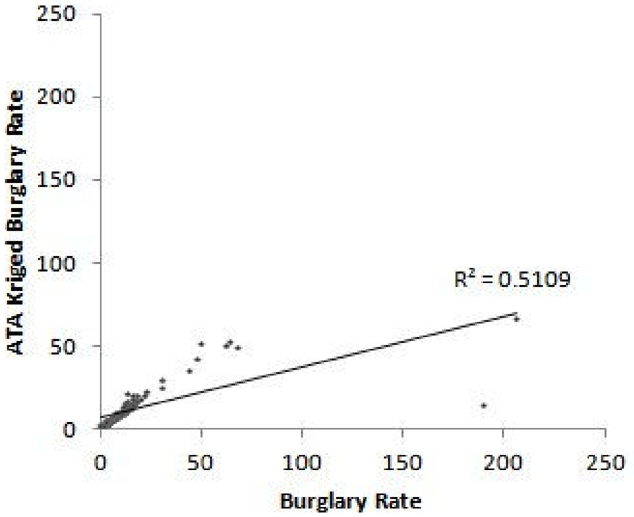


Figure 12: Correlation between ATA Poisson Kriged and Original Burglary Rates

ATP Poisson kriging was carried out using the same variogram and deconvolution models as Poisson kriging. ATP kriging estimates are presented in Figure 13, which shows a similar pattern to the ATA results. Locations with high ATP kriging estimates include Sunnydale to the south, the area around Islais Creek, Hunter's Point and the South-of-Market neighbourhood.

ATP kriging, like other types of kriging, may generate negative kriging estimates because kriging weights can be negative. In the case of this study, ATP kriging produced 414 negative estimates equalling 2.9% of the total number of points. The lowest kriging estimate was -10.323. Negative kriging estimates were corrected to 0, as suggested by Dr. Pierre Goovaerts (personal communication), because they constituted a low percentage of the total number of estimates. A map highlighting these adjusted values can be found in Appendix A (Figure A- 3).

The highest rates estimated by ATP kriging are higher than ATA estimates because they represent smaller areas. However, the results are coherent (before the adjustment of negative values to 0), and the mean of the ATP estimates within a polygon equals the ATA estimate for the area.

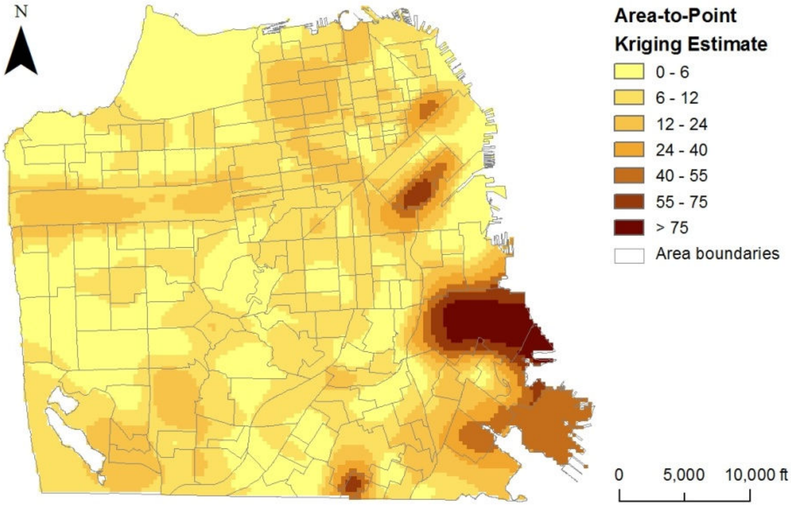


Figure 13: Burglary Rate Area-to-Point Kriging Estimate

ATP kriging variance is presented in Figure 14. In line with previous findings, the map shows similarities to the ATA kriging results. Kriging variance is highest in the Presidio, the Golden Gate Park, Hunter's Point and the area near Islais Creek. All of these neighbourhoods are located in large census tracts. Kriging variance is lowest in the downtown area to the north-east of the study area, where the polygons are smallest.

A summary of the kriging estimates for all types of kriging is presented in Table 3. It is important to note that for most datasets, the variance exceeds the mean. This violates the main assumption of the Poisson distribution; that variance and mean are equal. As previously discussed, the mean of the point estimates tends to be higher than that of the areas. Variance is highest for the observed rates and the ATP kriging results.

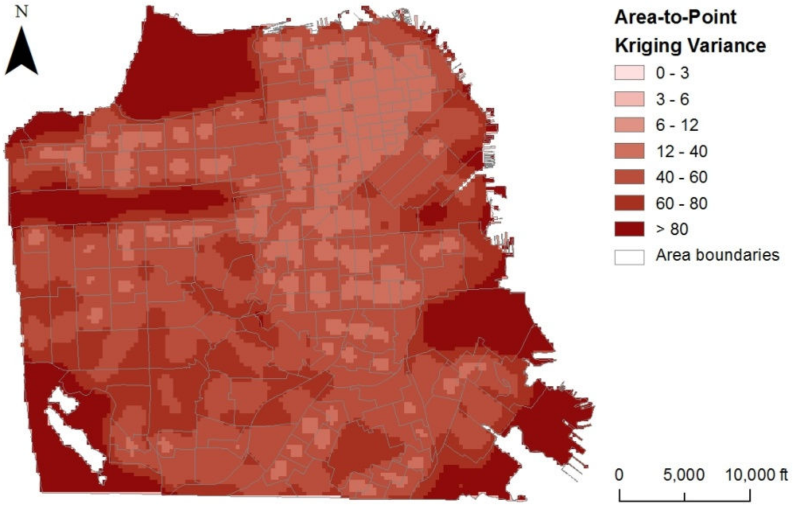


Figure 14: Burglary Rate Area-to-Point Kriging Variance

Table 3: Summary of Kriging Estimates

	Mean	Variance	Min	Max
Census tracts				
Observed rates	11.910	451.262	0	206.250
ATA Poisson kriging	10.329	81.846	1.674	66.809
Points				
Centroid method	12.294	84.909	3.364	56.058
ATP Poisson kriging	12.760	248.341	-10.323	106.278
Adjusted ATP Poisson kriging	12.826	246.255	0.000	106.278
Aggregates of point estimates[a]				
Centroid method	10.716	6.186	4.282	45.104
ATP Poisson kriging	10.315	81.489	2.082	70.268
Adjusted ATP Poisson kriging	10.343	81.316	2.170	70.268

[a] Aggregate of centroids of the point grid cells falling within the area boundaries

4.3. Local Indicators of Spatial Association

Local Indicators of Spatial Association, or Local Moran's *I* analysis, was carried out in order to search for local clusters or outliers in the datasets. It was performed using the original observed burglary rates in combination with the results of the three kriging methods.

Figure 15 presents the results of Local Moran's *I* analysis for the observed data. The Golden Gate Park is found to be a high-low outlier. It is an area with a high rate, surrounded by those with low rates. The area around Islais Creek and Hunter's Point is identified as a high-high cluster. This is a polygon with a high burglary rate, surrounded by those also with a high rate. All other census tracts have non-significant results.

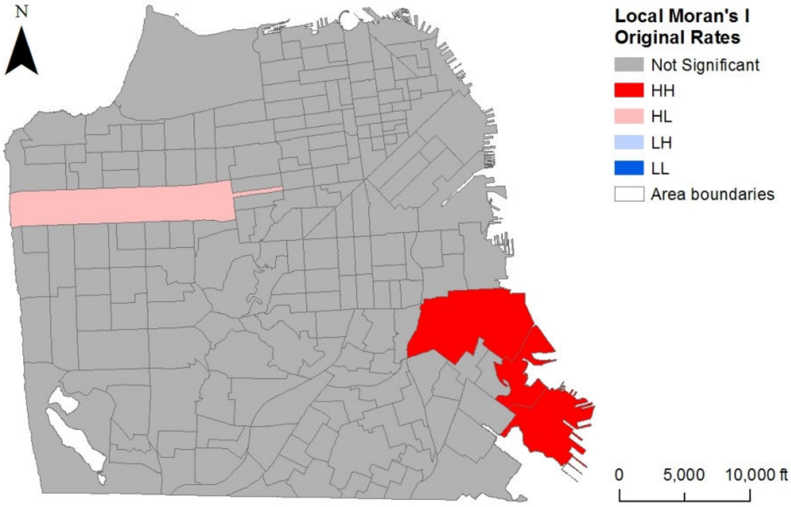


Figure 15: Local Moran's I, Original Burglary Rates

Local Moran's *I* analysis carried out on the ATA kriging estimates (Figure 16) demonstrates a different pattern. For example, one area near downtown San Francisco is a high-low outlier. The high-high cluster in the south-east now covers a far larger area. Sunnydale and the area to the north are also high-high clusters.

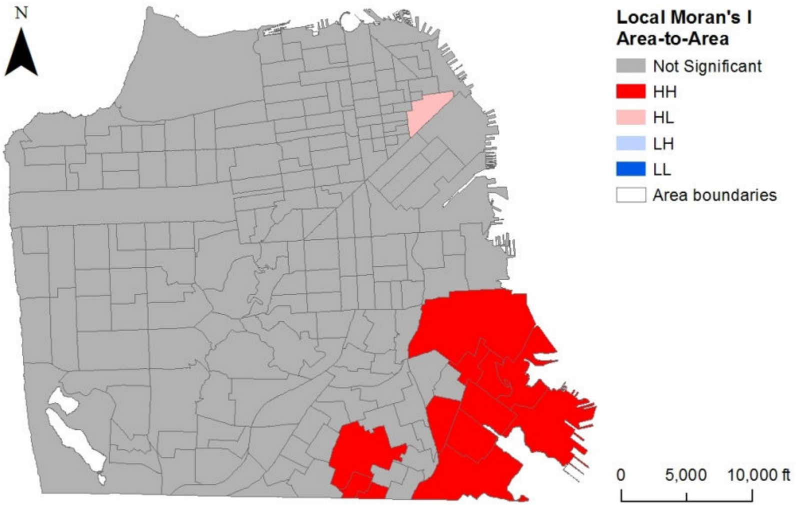


Figure 16: Local Moran's I, Area-to-Area Kriging Estimates

Figure 17 presents the results of the Local Moran's *I* analysis on the ATP kriging estimates. Whilst there are no outliers, there are many high-high and low-low clusters (areas with low values surrounded by those also with low values). Similarly to the Local Moran's *I* results for ATA kriging and the observed rates, there are high-high clusters or hotspots in Hunter's Point, the area around Islais Creek, Sunnydale, South-of-Market and downtown. The largest low-low cluster is located in the Presidio, with other clusters in the east of the study area by Mission Bay.

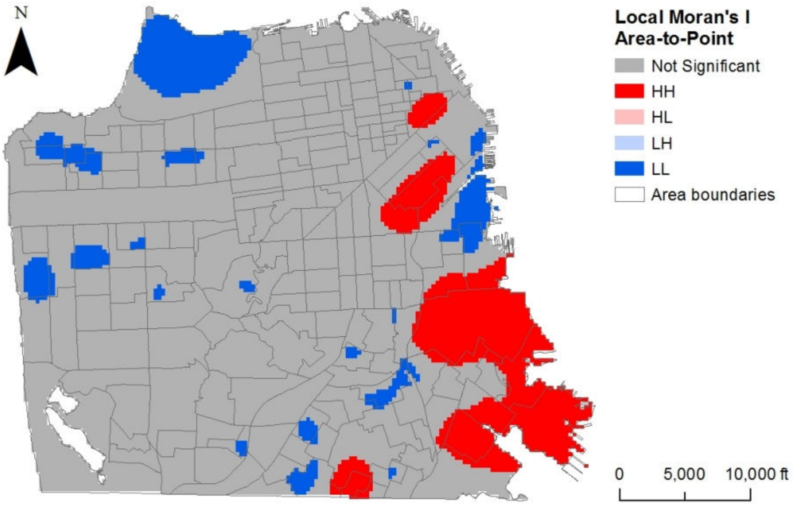


Figure 17: Local Moran's I, Area-to-Point Kriging Estimates

The results of the Local Moran's *I* analysis of the centroid method Poisson kriging estimates (Figure 18) depict a similar circular pattern to the kriging estimates themselves. The high-high cluster in the centre of the Golden Gate Park is located in the middle of low-low clusters. High-high clusters are also located in the south east area (Islais Creek and Hunter's Point), Sunnydale and the South-of-Market neighbourhood.

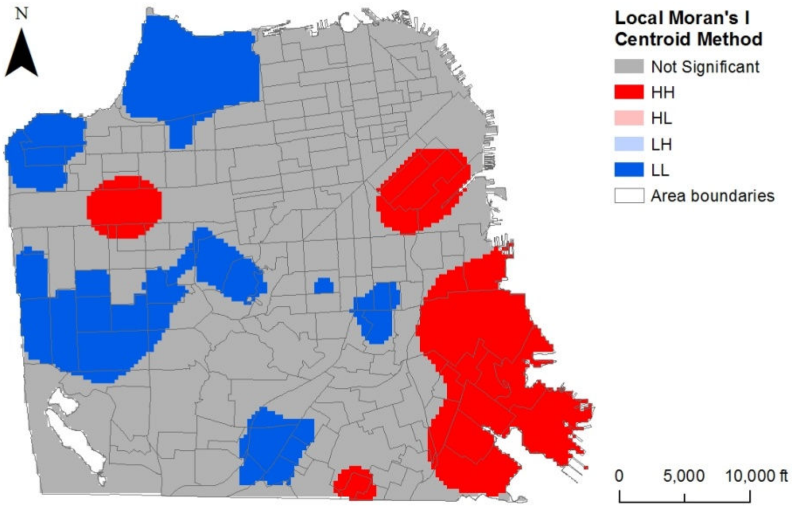


Figure 18: Local Moran's I, Centroid Method Kriging Estimates

A summary of the results of all Local Moran's *I* analyses is presented in Table 4. It is notable that there are no low-high clusters. It can be seen in the table that the vast majority of areas and points have no significant Moran's *I* value. The centroid method results in the highest percentages of both high-high and low-low clusters.

Table 4: Summary of Local Moran's *I*

	Percentage of total area within each category (numbers in brackets are the percentages of area counts)			
	Original rates	ATA	ATP	Centroid method
HH (High-High)	5.98 (1.55)	12.28 (6.19)	11.35	16.20
HL (High-Low)	3.75 (0.52)	0.47 (0.52)	0	0
LH (Low-High)	0	0	0	0
LL (Low-Low)	0	0	7.49	18.53
Not significant	90.26 (97.94)	87.26 (93.30)	81.16	65.27

4.4. Areal regression

The areas highlighted by the Local Moran's *I* analyses as residential burglary hotspots include Sunnydale, the area around Islais Creek, Hunter's Point and the South-of-Market neighbourhood. In order to investigate the relationship between burglary rates and the characteristics of localities, spatial and non-spatial regression analyses were conducted.

The five explanatory variables used in the regression analyses are detailed in Table 2. Of the 194 census tracts, 4 were excluded from the analysis due to missing data for one of the explanatory variables (house value). During the variable selection and regression modelling processes, the Variance Indicator Factor was calculated to quantify multicollinearity. All results were low (below 5), and therefore all datasets were included in the regression. The R code used for the analyses can be found in Appendix B.

4.4.1. Regression of original data

Poisson Regression was carried out using the original data for response and explanatory variables. The results can be found in Table 5. All explanatory variables, apart from education, have statistically significant results. A change of one unit in an explanatory variable leads to an expected change (equal to the parameter estimate) in the natural log of the response variable, when all other variables are held constant. For example, an increase of one unit in the drug incident rate is associated with an increase of 0.0124 in the natural log of the burglary rate. The regression parameters suggest that higher home values are related to lower burglary rates. This may imply that offenders commit burglaries near to their residences, and do not tend to travel to wealthier areas of the city. Areas with dense housing are linked to lower burglary rates, as was expected following the review of the literature. Relevant studies also suggested that areas with high drug use would experience higher drug rates. This is corroborated by the Poisson regression results. Higher median incomes are associated with slightly higher burglary rates, although the parameter estimate is very small. None of the parameter estimates are large. This effect is amplified by the use of the natural log link.

Table 5: Results of the Poisson Regression of original data

ln(Burglary Rate)	Poisson Regression			
Parameter	Estimate	Std. Error	z-value	Pr(>|z|)
(Intercept)	2.5206	0.121	20.803	< 0.001
Drug Rate	0.0124	0.000	26.360	< 0.001
Income	0.0081	0.001	7.924	< 0.001
Housing Density	-0.0016	0.000	-3.490	< 0.001
Education	-0.0022	0.003	-0.809	0.419
Home Value	-0.0102	0.001	-8.287	< 0.001
Null deviance	1940.8			
Residual deviance	1378.0			
RSS	34842.91			
Dean's Overdispersion Test	124.92, p-value < 0.001			

Figure 19 presents the burglary rates predicted by the fitted Poisson Regression model. The pattern broadly represents that of the original data, although there are fewer low values. A lot of deviance remains unexplained by the model. The Residual Sum of Squares (RSS) is large, with a value of 34842.91. A map of the residuals can be found in Appendix A (Figure A- 19). This map shows a smoothing of predicted values.

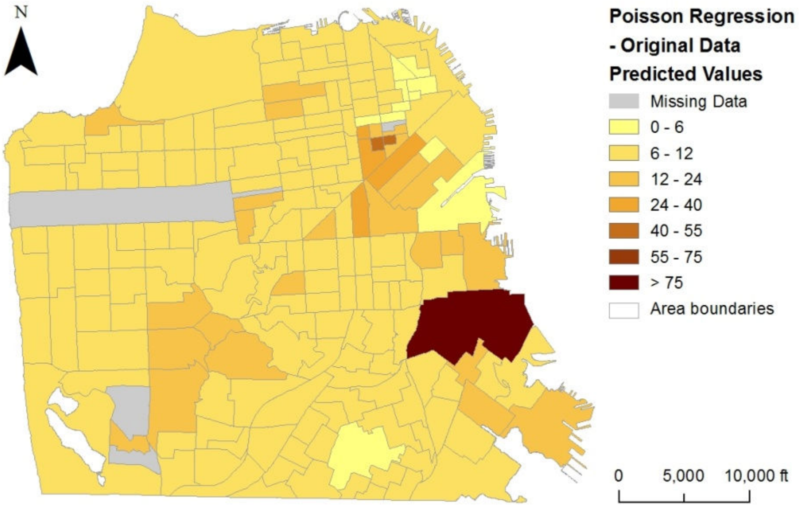


Figure 19: Poisson Regression: Original Data - Predicted Values

The results of Dean's overdispersion test indicate that there is evidence of overdispersion in the residuals of the Poisson regression model. To address this issue, Negative Binomial regression was carried out. The results can be seen in Table 6. For the Negative Binomial regression model, only the drug incident rate and home value are statistically significant. The directions of these relationships are the same as with the Poisson regression.

The residual deviance is a lot lower than for the Poisson regression, but the RSS is larger. The residual map can be found in Appendix A (Figure A- 20). Cragg and Uhler's pseudo R-squared value, a measure of goodness-of-fit that ranges from 0 to 1, is 0.503. This value is low, and indicates that the model parameters do not improve much upon the prediction of the null model (a model predicting the response variable without any explanatory variables).

Figure 20 presents the predicted values of the Negative Binomial regression model. The map shows an area of high values in the downtown area of San Francisco, and in the neighbourhood around Islais Creek.

Table 6: Results of the Negative Binomial Regression of original data

ln(Burglary Rate)	Negative Binomial Regression			
Parameter	Estimate	Std. Error	z-value	Pr(>|z|)
(Intercept)	3.2921	0.317	10.397	< 0.001
Drug Rate	0.0125	0.002	6.819	< 0.001
Income	0.0007	0.003	0.263	0.7927
Housing Density	-0.0008	0.001	-0.801	0.4233
Education	-0.0051	0.006	-0.869	0.3847
Home Value	-0.0127	0.003	-3.808	< 0.001
Null deviance	262.59			
Residual deviance	189.20			
Cragg and Uhler's pseudo r-squared	0.503			
RSS	55041.48			

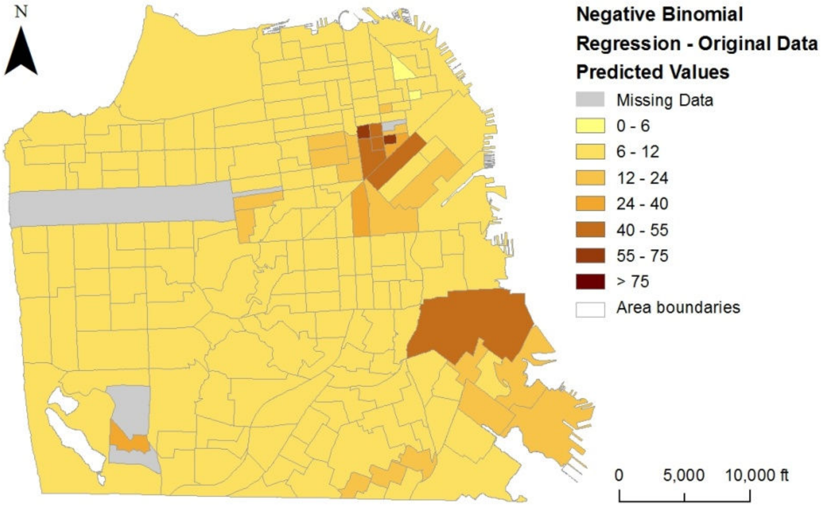


Figure 20: Negative Binomial Regression: Original Data - Predicted Values

4.4.2. Regression of ATA kriged data

For each of the ATA kriged explanatory variables used in this analysis, the maps, variogram models and deconvolution models are presented in Figures A-4 to A-18 in Appendix A.

The results of Poisson regression using the ATA kriged data are shown in Table 7. Only three of the parameter estimates are significant; drug incident rate, home value and education. The higher the percentage of over 25-year-olds with less than a high school education, the higher the burglary rate. Drug rate and home value have the same direction of relationship to burglary rate as in the previous regression analyses.

Table 7: Results of the Poisson Regression of ATA kriged data

ln(ATA Kriged Burglary Rate)	Poisson Regression			
Parameter	Estimate	Std. Error	z-value	Pr(>|z|)
(Intercept)	3.2568	0.193	16.918	< 0.001
Drug Rate	0.0027	0.001	3.602	< 0.001
Income	0.0016	0.001	1.182	0.237
Housing Density	0.0001	0.001	0.257	0.797
Education	0.0068	0.003	2.029	0.042
Home Value	-0.0144	0.002	-7.223	< 0.001
Null deviance	981.25			
Residual deviance	850.37			
RSS	14978.41			
Dean's Overdispersion Test	62.2048, p-value < 0.001			

The difference between null and residual deviance is low, suggesting the model does not extensively improve upon the predictions of the null model. The RSS is also large. A map of residuals can be found in Figure A- 21. The better fit for the regression model with ATA kriged data may be partly due to working with noise-filtered dependent and independent variables, which have lower variances.

The predicted values map (Figure 21) demonstrates the extent of the smoothing effect of the Poisson regression on ATA variables, which are themselves smoothed. There are fewer higher and low values, and most are within the 6 - 12 range.

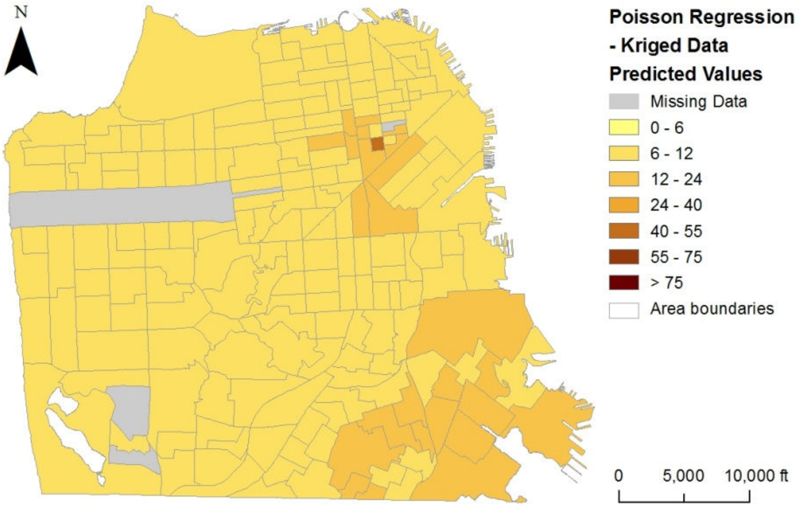


Figure 21: Poisson Regression: ATA Kriged Data - Predicted Values

The results of Dean's Overdispersion Test indicate that there is overdispersion in the residuals of the Poisson regression on ATA kriged data. This overdispersion is less than in the residuals of the Poisson regression of the original data, as would be expected due to the smoothing of all ATA kriged variables.

As a subsequent step, Negative Binomial regression of ATA kriged data was carried out due to this evidence of overdispersion. The results can be seen in Table 8. Similarly to the Negative Binomial regression of the original data, the only significant parameters are drug incident rate and home value, which have the same direction of relationship as in the previous analysis.

The values predicted by the Negative Binomial Regression model of ATA kriged data are presented in Figure 22. In line with previous findings, the smoothing effect of the model is evident in this map and also in the residual map (sees Figure A- 22). One notable high predicted value exists in the downtown area. As before, the majority of the predicted values are within the range of 6 to 12.

Table 8: Results of the Negative Binomial Regression of ATA kriged data

ln(ATA Kriged Burglary Rate)	Negative Binomial Regression			
Parameter	Estimate	Std. Error	z-value	Pr(>|z|)
(Intercept)	3.8315	0.406	9.446	< 0.001
Drug Rate	0.0047	0.002	2.805	0.005
Income	0.0017	0.003	0.628	0.530
Housing Density	0.0004	0.001	0.372	0.710
Education	0.0005	0.007	0.082	0.935
Home Value	-0.0209	0.004	-4.799	< 0.001
Null deviance	222.24			
Residual deviance	179.13			
Cragg and Uhler's pseudo R-squared	0.290			
RSS	19530.30			

The RSS is higher than for the Poisson regression model of ATA kriged data. The pseudo R-squared value for this model is only 0.290, which is even lower than for the Negative Binomial regression of the original data. This indicates that the model parameters improve little upon the prediction of the null model.

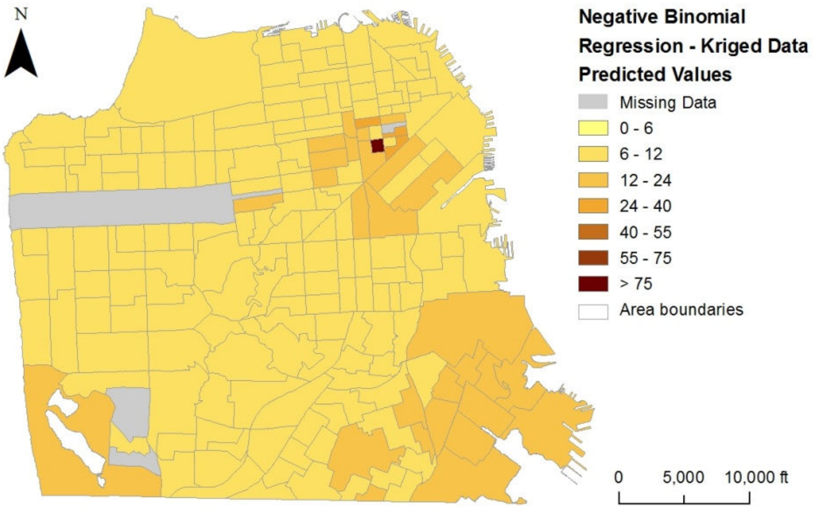


Figure 22: Negative Binomial Regression: ATA Kriged Data - Predicted Values

4.5. Geographically Weighted Regression

Geographically Weighted Poisson Regression (GWPR) was carried out on the original and ATA kriged response and explanatory variables. GWPR is often carried out when there is evidence of spatial autocorrelation in the residuals of regression. For the Poisson regression models with both types of data, the Global Moran's *I* values of the regression residuals are not statistically significant. However GWPR was undertaken in order to investigate local relationships between the explanatory and response variables.

GWPR was carried out using the housing-unit weighted centroids. Adaptive kernels were used for the weighting. The kernel bandwidths were chosen by minimising the AICc. For the original data the bandwidth was 36 neighbours, for the kriged data the bandwidth was 98. These were the bandwidths with the lowest AICc values that were computed without model convergence errors.

A summary of the minimum and maximum local parameter estimates for the GWPR regression of both types of data is presented in Table 9. All of the parameter estimates range from negative to positive relationships with the response variables, as parameter estimates are allowed to vary locally.

The difference in bandwidth is the likely reason why the RSS for the local kriged data model is higher than that of the original data model. The opposite would be expected, due to the results of the global models.

The values predicted by GWPR for the original data are presented in Figure 23. The results resemble the original rates more than any of the non-geographical regression models.

Table 9: Geographically Weighted Poisson Regression Results

	Original Data		ATA Kriged Data	
	Minimum	Maximum	Minimum	Maximum
Intercept	-1.404	6.130	1.174	6.925
Drug Rate	-0.075	0.064	-0.002	0.055
Income	-0.010	0.011	-0.018	0.013
Housing Density	-0.025	0.019	-0.012	0.011
Education	-0.044	0.048	-0.034	0.017
Home Value	-0.054	0.044	-0.053	0.003
RSS (Global Model)	34843		14978	
RSS (Local Model)	4055		7985	

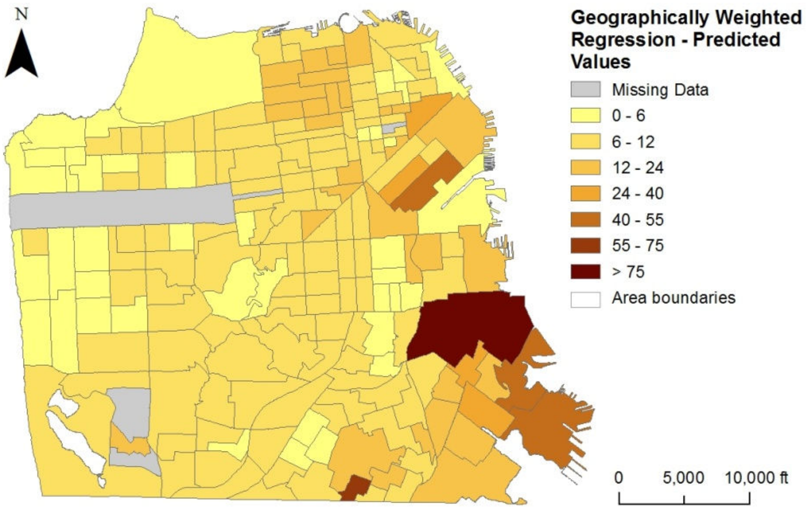


Figure 23: Geographically Weighted Regression: Original Data - Predicted Values

Local R-squared values for the GWPR model of the original data (Figure 24) show that the explanatory variables best explain burglary rates in the south-east of the study area. Local R-squared values are lowest to the south of the Golden Gate Park.

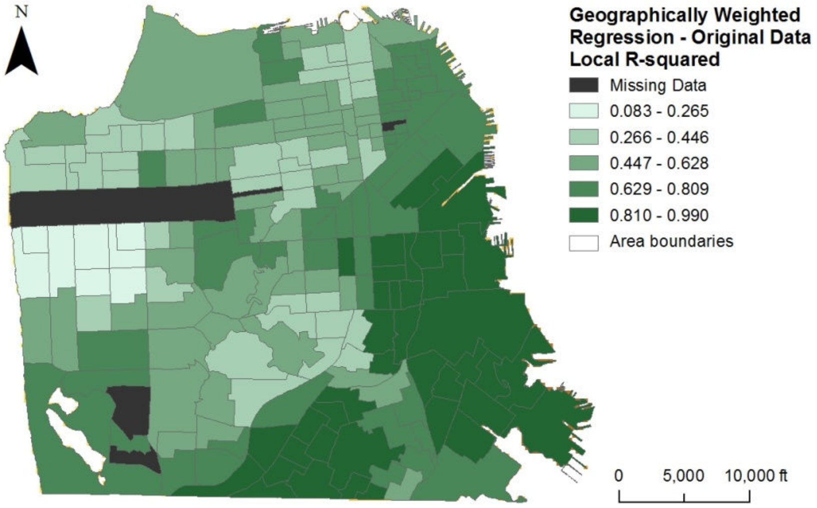


Figure 24: Geographically Weighted Regression: Original Data - Local R-squared

Predicted values for the GWPR of the ATA kriged data (Figure 25) do not visually closely resemble the maps of the response variable. Indications of smoothing exist, with high and low values being made less extreme. Despite this finding, the pattern of high ATA kriged rates in the south-east of the study area remains.

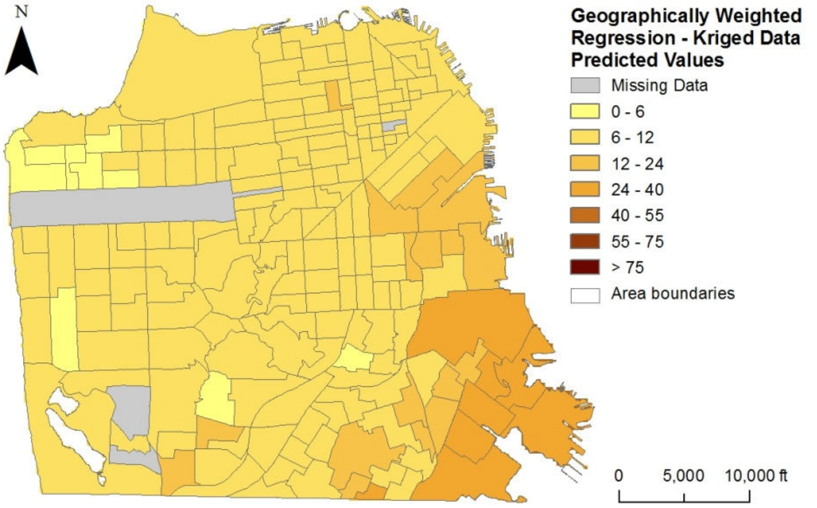


Figure 25: Geographically Weighted Regression: ATA Kriged Data – Predicted Values

Figure 26 displays the local R-squared values for the GWPR model for the ATA kriged datasets. The values are lower than those in the GWPR model for the original datasets. This suggests that the original dataset model explains variation in the original burglary rates more sufficiently than the ATA kriged dataset model explains variation in ATA kriged burglary rates. This may be due to the effect of the smoothed data, or due to the larger adaptive kernel size used for the GWPR of ATA kriged data. However, the models are not directly comparable due to the difference in response variables.

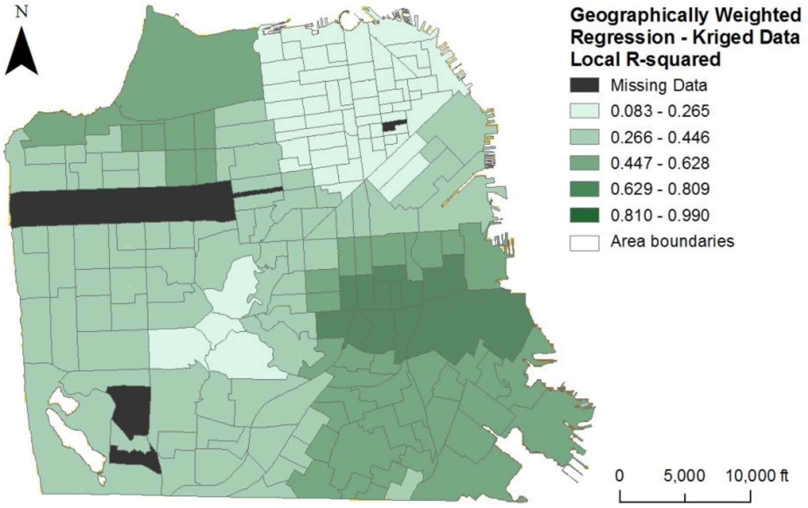


Figure 26: Geographically Weighted Regression: ATA Kriged Data - Local R-squared

5. CONCLUSIONS

The overall aim of this study was to investigate statistical approaches to analysing and visualising areal crime data. This was done through a case study of residential burglary in San Francisco, in the United States. The application of ATA and ATP Poisson Kriging, new approaches to the interpolation of areal count or rate data, have proven to be promising alternatives to the traditional method of point kriging from geographical centroids. Both the original and interpolated data were used as inputs into further analyses of spatial clusters and the relationship between socio-economic characteristics and burglary rates. This was done in order to explore how the use of interpolated data affects the results of such analyses.

The first objective of this study was to compare Area-to-Area and Area-to-Point kriging to Choropleth mapping and the traditional centroid method for interpolating residential burglary rates. Based on the results of this study it is evident that the centroid method has several limitations. Among these is the fact that concentric circles are clearly visible in the kriging estimates. This is unlikely to reflect reality due to spatial features that tend to delimit areas of high crime, such as main roads or changes of land use. The existence of irregularly shaped or large polygons also causes problems. In particular, the limitations of this method stem from its main assumption. It is unlikely that all crimes occur at the geographical centroid. An improvement on this method would be to use housing-unit-weighted centroids.

In comparison to the centroid method, Area-to-Area and Area-to-Point Poisson kriging may offer an improvement by utilising housing-unit-weighted centroids and a housing-unit grid cell map. In this study, ATA kriging estimates smooth the burglary rates and also reduce the influence of rates in areas with small populations. One possible limitation of the results is that there is low correlation between ATA kriging estimates and the original burglary rates. However, this is mostly caused by the smoothing of two large rates in the original dataset, both calculated from relatively small housing denominators. This method presents an obvious benefit of the ATA kriging method in comparison to Choropleth maps.

In particular, it is difficult to ascertain the quality of the burglary rate ATP kriging estimates without knowledge of the underlying risk of burglary. Observed burglary rates do not exist at the interpolation grid level, and observed counts of crime incidents per grid cell merely represent a realisation of the underlying risk of burglary. However, ATP kriging estimates decrease the visual bias caused by large areas as the result is a continuous surface. Similarly to the ATA kriging results, high rates calculated from small denominators are decreased. In comparison to the results from the centroid method of kriging, ATP kriging estimates display little evidence of concentric circles. Additionally, the influence of large or irregularly shaped polygons is diminished.

Based on the results of the analysis, it is apparent that for this study area, Area-to-Area and Area-to-Point Kriging improve on existing methods for presenting crime rate data such as Choropleth maps and Poisson kriging using centroids.

The second objective of the study was to locate spatial clusters of high or low crime rates in the study area. This was conducted using Local Moran's *I* tests on the original rates, and the ATA, ATP and centroid method kriging estimates. A review of the data shows that a number of areas of San Francisco are statistically significant high-high clusters of burglary rates in many of the tests. These neighbourhoods include Islais Creek, Hunter's Point and Sunnydale. Statistically significant low-low clusters were only found in the ATP and centroid method kriging estimates – those calculated from points instead of polygons. The only low-low cluster that appears on both maps is located in the Presidio. It has also become evident that the effect of using data kriged onto an interpolation grid is to produce more clusters. Following the conclusion of the first objective, that ATA kriging estimates are an improvement on Choropleth mapping, it could also be concluded that the Local Moran's *I* map produced with these estimates is an improvement on the map which uses the original rates.

The final objective of this study was to explore the relationship between residential burglary and socio-economic variables in the study area using both non-spatial and spatial regression techniques. This was achieved using the original data and the ATA kriged data as inputs for non-spatial Poisson and Negative Binomial Regression Models. Furthermore, both types of data were used as inputs into Geographically Weighted Poisson Regression. The use of ATA kriged data was intended to help alleviate the problems which may arise from performing regression analysis at scales that may misrepresent the relationship between response and explanatory variables.

The non-spatial regression results for the original data indicate that drug incident rate has a positive correlation with burglary rates, whilst home value has a negative correlation. However, the models do not represent a considerable improvement on the predictions of the null model, and the residuals are large. Non-spatial regression of ATA kriged data indicates that drug incident rate and home value have the same relationships to the response variable as the model using the original data. Residuals are lower than the original data model, but still relatively large. The results of the regression models using the ATA kriged data produce smoothed predicted values.

Geographically Weighted Poisson Regression improves on the fit of the Poisson Regression model. The results for the GWPR of the original and ATA kriged data show that all of the explanatory variables have both negative and positive correlations to the response variable, as parameter estimates vary over the study area. The predicted values for the GWPR of the original datasets closely resemble the original burglary rates. For the ATA kriged data, the predicted values are once again smoothed. Local R-squared values are higher for the model using original data than those of the model of ATA kriged data. Despite the improved fit of the GWPR, the residuals are still high.

The results of the regression analyses show that none of the spatial or non-spatial models adequately explain variance in burglary rates in San Francisco. However, the findings provide valuable insights into the neighbourhood characteristics that relate to high burglary rates. In particular, low home values and high drug incident rates may be associated with increased rates of residential burglary.

The techniques employed in this study build on the traditional methods for crime analysis and visualisation such as Choropleth mapping and non-spatial regression analyses. Based on the findings of this study, it can be stated that Area-to-Area and Area-to-Point kriging methods appear to improve on existing approaches to the interpolation of areal data. Visualisation is enhanced through the smoothing of rates based on small denominators. Visual bias is decreased when applying Area-to-Point kriging. Furthermore, use of the kriging estimates of these techniques as inputs into cluster and regression analyses provides an additional method which can be used to explore relationships at different scales. However, caution must be exercised when utilising these methods. There are some important limitations to the techniques used in this study, which are discussed in the next section.

5.1. Limitations

Whilst the techniques used in this analysis seem to offer improvements upon existing methods, care should be taken when utilising such an approach and interpreting the results. All of the methods, apart from the Local Moran's *I,* assume that the data has a particular distribution. None of the datasets used in this analysis have Poisson distributions, as the variances are higher than the means. Whilst rate or count data may have a Poisson distribution, this is not always the case. Crime incident data represents reality, and cannot be adjusted to the demands of the method.

A more thorough investigation into the suitability of ATA and ATP Poisson kriging methods would require knowledge of the underlying risk. Whilst point crime incident data is available in this case, this is merely a realisation of the underlying risk. A more complete evaluation could involve the use of simulated data.

An important consideration in the interpretation of this study is that ATA and ATP kriging cannot actually create higher resolution data from areas. Whilst such kriging methods can provide another useful visualisation and analysis technique, they are not a substitute for higher resolution data. The original data is subject to the MAUP, and therefore, the results of any analysis using this data will also have this limitation.

Similarly, it was suggested by Goovaerts (2006b) that using ATA kriged data may help alleviate problems caused by performing regression analysis at scales that may misrepresent the relationship between response and explanatory variables. This is unlikely to be the case, for the previously stated reason that ATA and ATP kriging cannot realistically be a replacement for data collected at different scales.

Another important limitation of ATP kriging is that in the case of this study, 2.9% of ATP estimates were negative. This problem may be due to overdispersion in the data or clustering of the areas. Obviously rates cannot be negative, and an improvement on the method would allow constraints to be placed on the kriging estimates.

There are two cases in the original data with very high burglary rates that may be seen as outliers. One of the cases with a high rate was excluded from the analysis due to missing data. The decision to not exclude the very high rates from the regression was taken because areas with high crime rates (hotspots) are important in policing. Knowledge of the socio-economic characteristics of neighbourhoods with very high burglary rates may be more important than understanding areas with more average rates. However, this means that the influence of the one remaining high value on the regression parameters is likely to be fairly high.

Another potential limitation of the regression analysis is that it investigates residential burglary based on the characteristics of the neighbourhoods of the victims, not the necessarily the neighbourhoods of the offenders. Whilst these may be the same areas, ideally data on the locations of offenders would also be analysed. This type of data is not available to the public for security reasons. As is typical with the regression analysis of crime data, this study is subject to the ecological fallacy - inferences about the actions of individuals are made from aggregated data.

A limitation of the study is that the log likelihood could not be calculated in R for the Poisson Regression models, and therefore measures of goodness of fit could also not be computed. This is because the response variable is non-integer. Such problems can be overcome by using the counts as the response variable and including an offset in the regression equation. However, this approach was not taken in this study because for the ATA kriged data, only rates are available.

Within the regression analyses, there exists another possible limitation. As stated in the literature review, Negative Binomial regression may still experience problems with overdispersion.

Another limitation is that one of the inputs of ATA and ATP kriging is a housing unit map, produced based on the assumption of homogenous housing unit density within census blocks. This assumption is unlikely to prove true. However, this is a commonly used technique and it was felt that this assumption was acceptable in the context of this study.

There has been criticism of GWR as a modelling tool. Any multicollinearity within the data may be increased by calculating local GWR coefficients. Therefore GWR should only be used as an explanatory tool, and the parameter estimates should be interpreted with caution.

5.2. Future work

This section briefly outlines some suggestions for future work on the topic of analysis and visualisation of areal crime data.

The results of ATP kriging could be simulated and used an input for LISA analysis. This would enable the testing of how uncertainty about the crime rates impacts the results the analysis. P-field simulation is the type of simulation most commonly used for simulating the results of ATP kriging (Goovaerts, 2006a). However, this method is unsuitable for the ATP kriged burglary rate dataset. The simulated values are unrealistic as there are too many negative results, probably caused by high kriging variance. An alternative approach to simulating the data would be to reverse the process, and simulate the original areal data using p-field simulation, and then perform ATP kriging. However, carrying out such an analysis within the time frame of this study was not possible.

Another approach that would be interesting to explore in future studies is Bayesian Hierarchical Modelling. This method is commonly used to visualise and analyse areal crime data. The study could also be expanded by carried out Geographically Weighted Regression on the ATP kriging results. Also of interest as an exploratory data analysis tool would be Geographically Weighted Negative Binomial Regression, although this method has not been implemented in any of the commonly used software packages.

REFERENCES

Almeida, E.S.d., Haddad, E.A. & Hewings, G.J.D., 2005. The Spatial Pattern of Crime in Minas Gerais: An Exploratory Analysis. *Economia Aplicada*, 9(1), pp.39-55.

Anselin, L., 1995. Local Indicators of Spatial Association - LISA. *Geographical Analysis*, 27(2), pp.93-115.

Anselin, L. et al., 2000. *Measurement and Analysis of Crime and Justice: Spatial Analyses of Crime*. Washington, D.C.: U.S. Department of Justice.

Berk, R. & MacDonald, J.M., 2008. Overdispersion and Poisson Regression. *Journal of Quantitative Criminology*, 24, pp.269-84.

Bernasco, W. & Luykw, F., 2003. Effects of Attractiveness Opportunity and Accessibility to Burglars on Residential Burglary Rates of Urban Neighborhoods. *Criminology*, 41(3), pp.981-1002.

Bowers, K. & Hirschfield, A., 2001. Introduction. In A. Hirschfield & K. Bowers, eds. *Mapping and Analysing Crime Data: Lessons from Research and Practice*. London: Taylor & Francis.

Brantingham, P. & Brantingham, P., 2008. Crime Pattern Theory. In R. Wortley & L. Mazerolle, eds. *Environmental Criminology and Crime Analysis*. Cullompton, UK: Willan Publishing. pp.78-95.

Breetzke, G.D., 2012. The effect of altitude and slope on the spatial patterning of burglary. *Applied Geography*, 34, pp.66-75.

California Penal Code Section 459, 2010.

Cragg, J.G. & Uhler, R., 1970. The demand for automobiles. *Canadian Journal of Economics*, 3(3), pp.386-406.

Cressie, N. & Read, T.R.C., 1989. Spatial Data Analysis of Regional Counts. *Biometrical Journal*, 81, pp.699-719.

Deadman, D., 2003. Forecasting residential burglary. *International Journal of Forecasting*, 19, pp.567-78.

Dean, C.B., 1992. Testing for overdispersion in Poisson and binomial regression models. *Journal of the American Statistical Association*, 87, pp.451-57.

Diehr, P., 1984. Small Area Statistics: Large Statistical Problems. *American Journal of Public Health*, 74(4), pp.313-14.

Duffala, D.C., 1976. Convenience Stores, Armed Robbery, and Physical Environmental Features. *American Behavioral Scientist*, 20, pp.227-45.

Eck, J.E. et al., 2005. *Mapping Crime: Understanding Hot Spots*. Washington, D.C.: U.S. Department of Justice.

Fotheringham, A.S., Charlton, M.E. & Brunsdon, C., 2002. *Geographically Weighted Regression. The analysis of Spatially Varying Relationships*. Chichester: Wiley.

Goovaerts, P., 2005. Geostatistical Analysis of Disease Data: Estimation of Cancer Mortality Risk From Empirical Frequencies using Poisson Kriging. *International Journal of Health Geographics*, 4, p.31.

Goovaerts, P., 2006a. Geostatistical Analysis of Disease Data: Visualization and Propagation of Spatial Uncertainty in Cancer Mortality Risk using Poisson Kriging and p-field Simulation. *International Journal of Health Geographics*, 5, p.7.

Goovaerts, P., 2006b. Geostatistical Analysis of Disease Data: Accounting for Spatial Support and Population Density in the Isopleth Mapping of Cancer Mortality Risk Using Area to Point Poisson Kriging. *International Journal of Health Geographics, 5*, p.52.

Goovaerts, P., 2008a. Kriging and Semivariogram Deconvolution in the Presence of Irregular Geographical Units. *Mathematical Geosciences*, 1(40), pp.101-28.

Goovaerts, P., 2008b. Geostatistical Analysis of Health Data: State-of-the-art and Perspectives. In A. Soares, M.J. Pereira & R. Dimitrakopoulos, eds. *GeoENV VI - Geostatistics for Environmental Applications: Proceedings of the Sixth European Conference on Geostatistics for Environmental Applications*. Netherlands: Springer. pp.3-22.

Goovaerts, P. & Jacquez, G.M., 2005. Detection of temporal changes in the spatial distribution of cancer rates using local Moran's I and geostatistically simulated spatial neutral models. *Journal of Geographical Systems*, 7(1), pp.137-59.

Gotway, C.A. & Wolfinger, R.D., 2003. Spatial prediction of counts and rates. *Statistics in Medicine*, 22, pp.1415-32.

Gotway, C.A. & Young, L.J., 2002. Combining Incompatible Spatial Data. *Journal of the American Statistical Association*, 97(458), pp.632-48.

Haining, R.P., Kerry, R. & Oliver, M.A., 2010. Geography, Spatial Data Analysis, and Geostatistics: An Overview. *Geographical Analysis*, 42, pp.7-31.

Ham-Rowbottom, K.A., Gifford, R. & Shaw, K.T., 1999. Defensible Space Theory and the Police: Assessing the Vulnerability of Residences to Burglary. *Journal of Environmental Psychology*, 19, pp.117-29.

Hearnden, I. & Magill, C., 2004. *Decision-making by house burglars: offenders' perspectives*. London: Home Office.

Johnson, G.D., 2004. Small area mapping of prostate cancer incidence in New York State (USA) using fully Bayesian hierarchical modelling. *International Journal of Health Geographics*, 3, p.29.

Kerry, R., Goovaerts, P., Haining, R.P. & Ceccato, V., 2010. Applying Geostatistical Analysis to Crime Data: Car-Related Thefts in the Baltic States. *Geographical Analysis, 42*, pp.53-77.

Krivoruchko, K., Gotway, C.A. & Zhigimont, A., 2003. Statistical Tools for Regional Data Analysis Using GIS. In *GIS'03, Proceedings of the 11th ACM*

International Symposium on Advances in Geographical Information Systems. New Orleans, Louisiana, 2003. ACM, New York.

Kyriakidis, P.C., 2004. A Geostatistical Framework For Area-To-Point Spatial Interpolation. *Geographical Analysis, 36*, pp.259-89.

La Vigne, N.G. & Groff, E.R., 2001. The evolution of crime mapping in the United States: from the descriptive to the analytic. In A. Hirschfield & K. Bowers, eds. *Mapping and Analysing Crime Data: Lessons from Research and Practice*. London: Taylor & Francis.

Laukkanen, M., Santtila, P., Jern, P. & Sandnabba, K., 2008. Predicting offender home location in urban burglary series. *Forensic Science International*, 176, pp.224-35.

Law, J. & Haining, R., 2004. A Bayesian Approach to Modeling Binary Data: The Case of High-Intensity Crime Areas. *Geographical Analysis*, 36, pp.197-216.

Malczewski, J. & Poetz, A., 2005. Residential Burglaries and Neighborhood Socioeconomic Context in London, Ontario: Global and Local Regression Analysis. *The Professional Geographer*, 57(4), pp.516-29.

Males, M.A., 2009. *San Francisco perpetuates "dark ages" crime prejudices*. [Online] Available at: http://www.cjcj.org/post/juvenile/justice/san/francisco/perpetuates/dark/ages/crime/prejudices/0 [Accessed 28 September 2011].

Mullins, C.W. & Wright, R., 2003. Gender, Social Networks and Residential Burglary. *Criminology*, 41(3), pp.813-40.

Nee, C. & Meenaghan, A., 2006. Expert Decision-Making in Burglars. *British Journal of Criminology*, 46, pp.935-49.

Openshaw, S. & Taylor, P., 1979. A million or so correlation coefficients: three experiments on the modifiable area unit problem. In N. Wrigley, ed. *Statistical Applications in the Spatial Sciences*. London: Pion. p.127-144.

Osgood, W.D., 2000. Poisson-based Regression Analysis of Aggregate Crime Rates. *Journal of Quantitative Criminology, 16*, pp.21-43.

Pease, K., 2001. Decision support in crime prevention: data analysis, policy evaluation and GIS. In A. Hirschfield & K. Bowers, eds. *Mapping and Analysing Crime Data: Lessons from Research and Practice*. London: Taylor & Francis.

Ratcliffe, J.H., 2002. Damned If You Don't, Damned If You Do: Crime Mapping And Its Implications In The Real World. *Policing and Society*, 12(3), pp.211-25.

Sherman, L.W., Gartin, P.R. & Buerger, M.E., 1989. Hot spots of predatory crime: Routine activities and the criminology of place. *Criminology*, 27, pp.27-55.

State of California, 2011. *California's unemployment rate decreases to 11.7 percent*. [Online] Available at: http://www.edd.ca.gov/About_EDD/pdf/urate201111.pdf.

Sugiura, N., 1978. Further analysis of the data by akaike's information criterion and the finite corrections. *Communications in Statistics - Theory and Methods*, 7(1), pp.13-26.

Tiefelsdorf, M. & Wheeler, D., 2005. Multicollinearity and correlation among local regression coefficients in geographically weighted regression. *Journal of Geographical Systems*, 7(2), pp.161-87.

Turton, I. & Openshaw, S., 2001. Methods for automating the geographical analysis of crime incident data. In A. Hirschfield & K. Bowers, eds. *Mapping and Analysing Crime Data: Lessons from Research and Practice*. London: Taylor & Francis.

U.S Census Bureau, 2000. *Decennial Management Division Glossary*. [Online] Available at: http://www.census.gov/dmd/www/glossary.html [Accessed 15 September 2011].

U.S. Census Bureau, 2010a. *American FactFinder, U.S. Census 2010*. [Online] Available at: http://factfinder2.census.gov/.

U.S. Census Bureau, 2010b. *American FactFinder: Profile of General Population and Housing Characteristics*. [Online] Available at: http://factfinder2.census.gov/faces/tableservices/jsf/pages/productview.xhtml?pid=DEC_10_DP_DPDP1&prodType=table.

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

Wallace, A., 2009. Mapping City Crime and the New Aesthetic of Danger. *Journal of Visual Culture*, 8(1), pp.5-24.

Williamson, D. et al., 2001. Tools in the spatial analysis of crime. In A. Hirschfield & K. Bowers, eds. *Mapping and Analysing Crime Data: Lessons from Research and Practice*. London: Taylor & Francis.

Yoo, E.-H., Kyriakidis, P.C. & Tobler, W., 2010. Reconstructing Population Density Surfaces from Areal Data: A Comparison of Tobler's Pycnophylactic Interpolation Method and Area-to-Point Kriging. *Geographical Analysis*, 42, pp.78-98.

Zhu, L., Gorman, D.M. & Horel, S., 2006. Hierarchical Bayesian spatial models for alcohol availability, drug "hot spots" and violent crime. *International Journal of Health Geographics*, 5, p.54.

APPENDICES

APPENDIX A

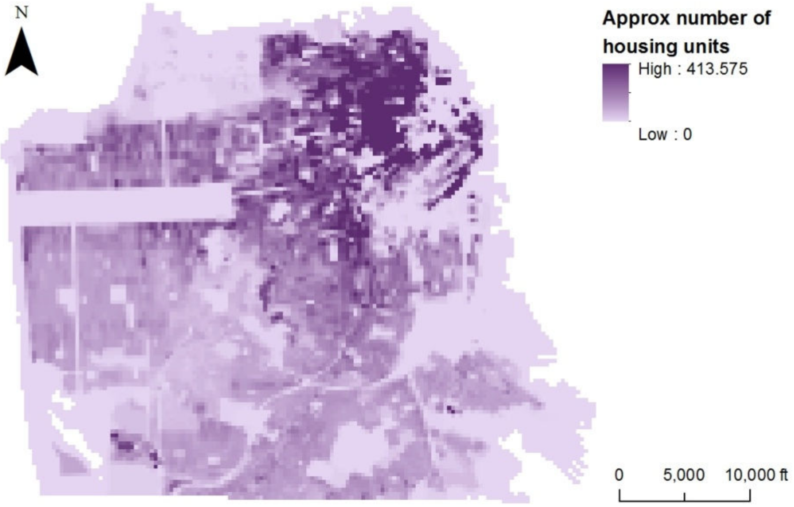


Figure A- 1: Raster map of the approximate number of housing units per grid cell

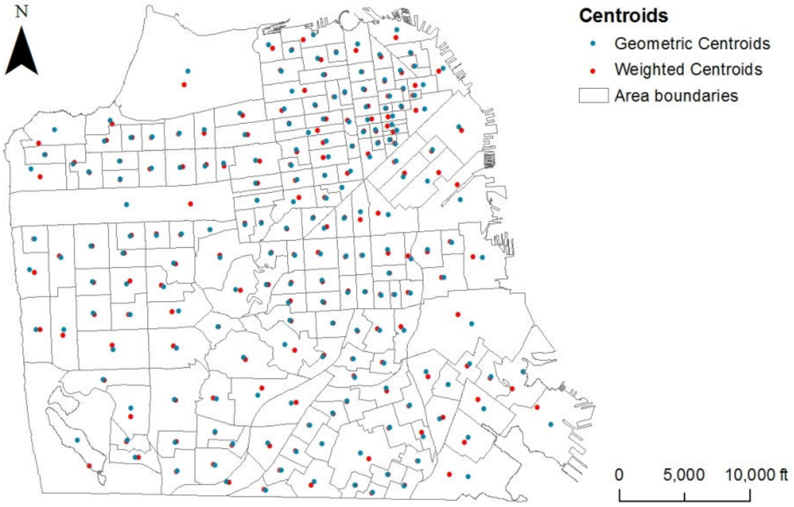


Figure A- 2: Geometric and Housing Unit Weighted Centroids

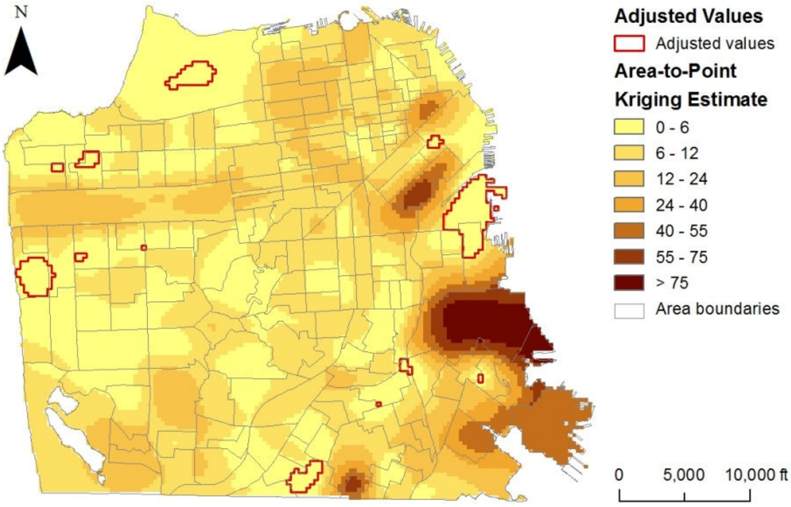


Figure A- 3: Burglary Rate Area-to-Point Kriging Estimates adjusted from negative values to 0

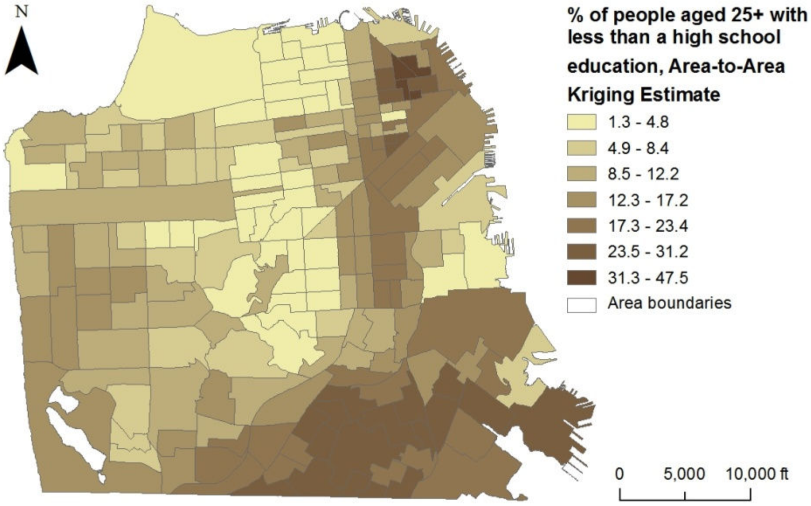


Figure A- 4: Percentage of population aged 25 or over with less than a high school education, Area-to-Area Kriging Estimates

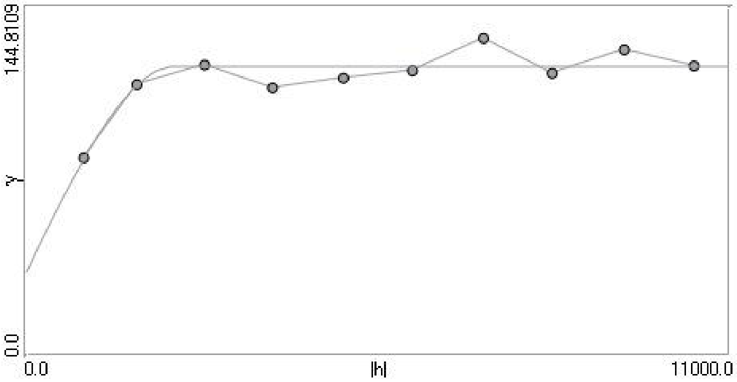


Figure A- 5: Education Variogram

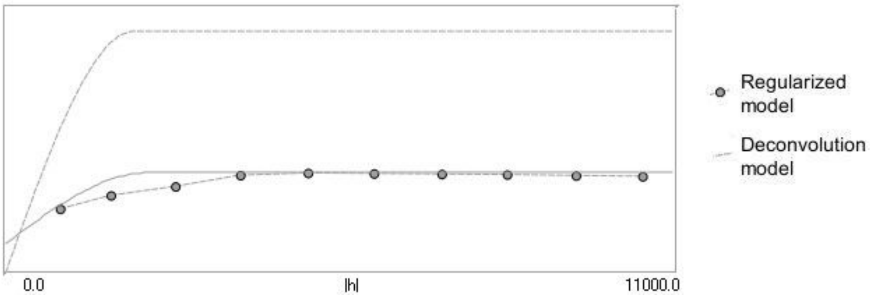


Figure A- 6: Education Deconvolution Model

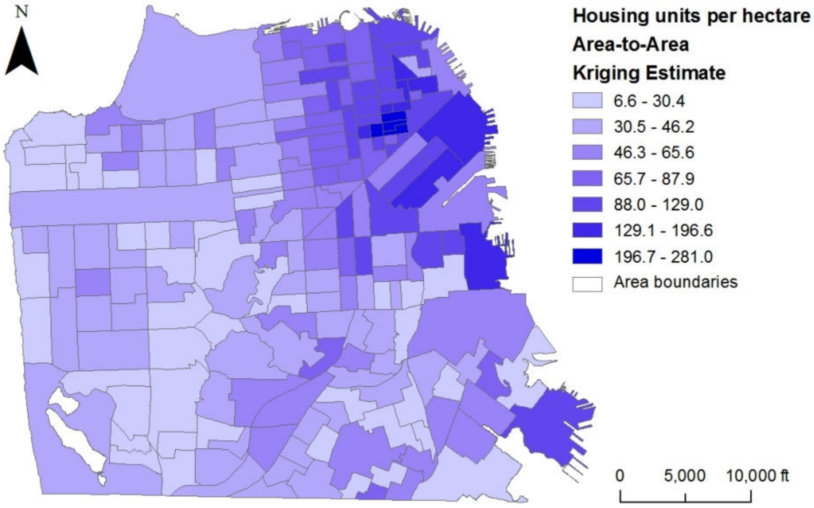


Figure A- 7: Housing Units per Hectare, Area-to-Area Kriging Estimates

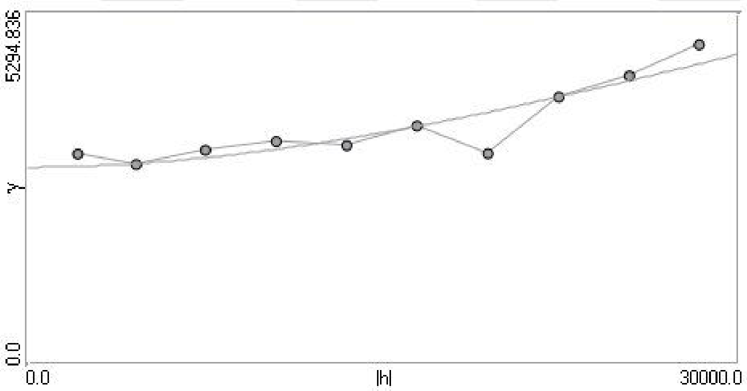


Figure A- 8: Housing Density Variogram

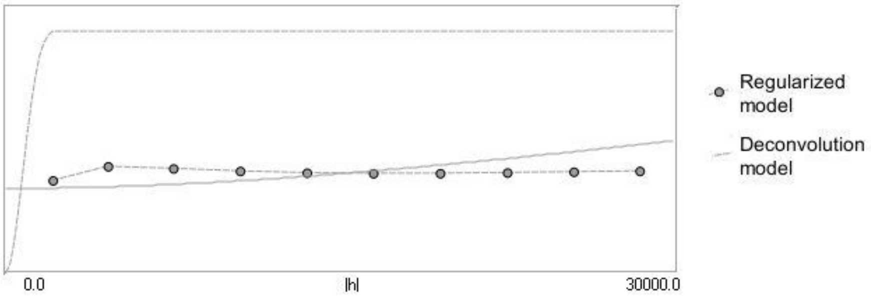


Figure A- 9: Housing Density Deconvolution Model

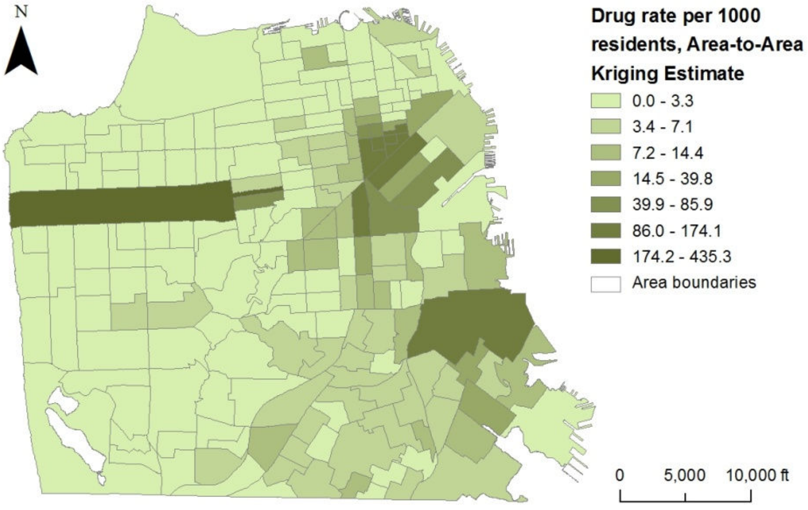


Figure A- 10: Drug Incidents per 1000 residents, Area-to-Area Kriging Estimates

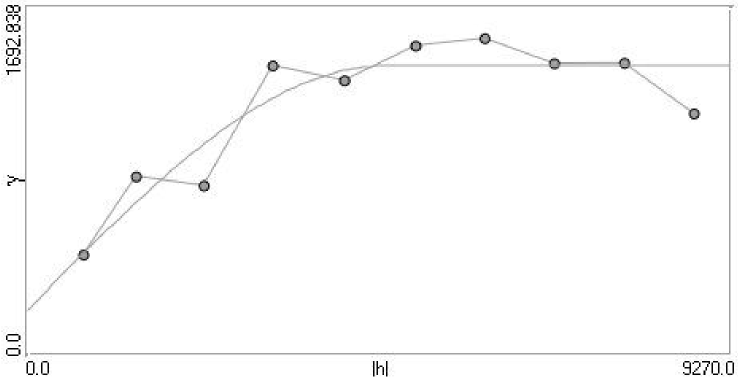


Figure A- 11: Drug Rate Variogram

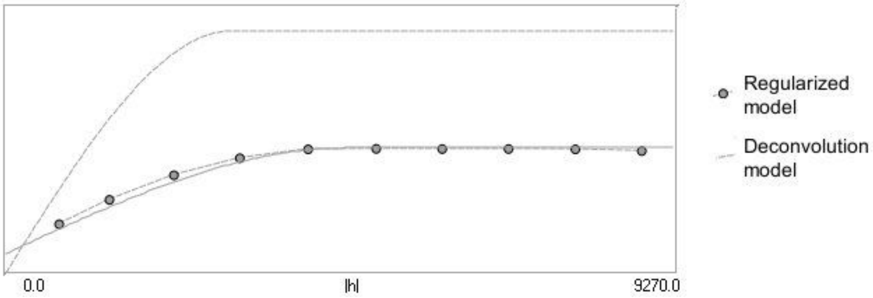


Figure A- 12: Drug Rate Deconvolution Model

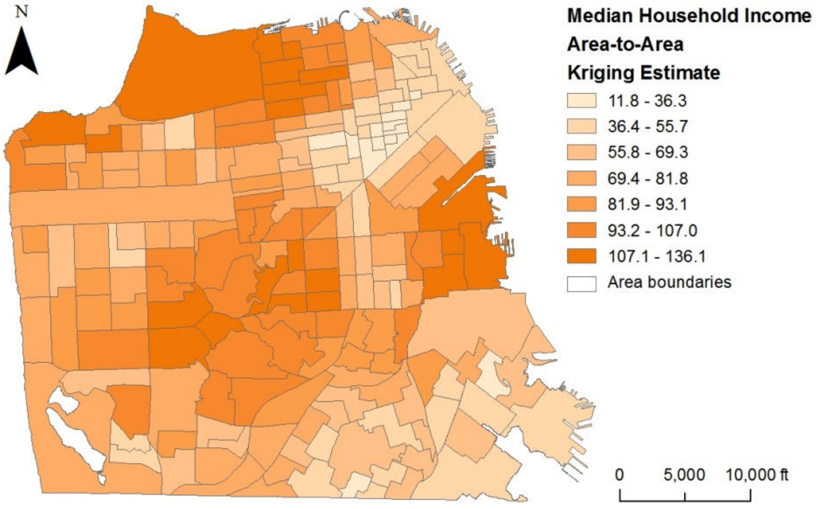


Figure A- 13: Median Household Income, Area-to-Area Kriging Estimates

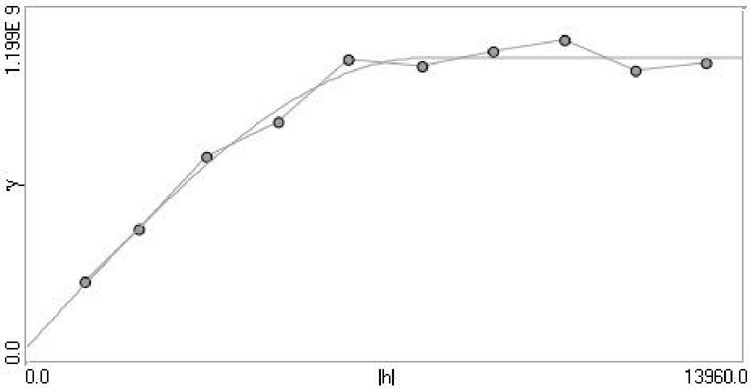


Figure A- 14: Income Variogram

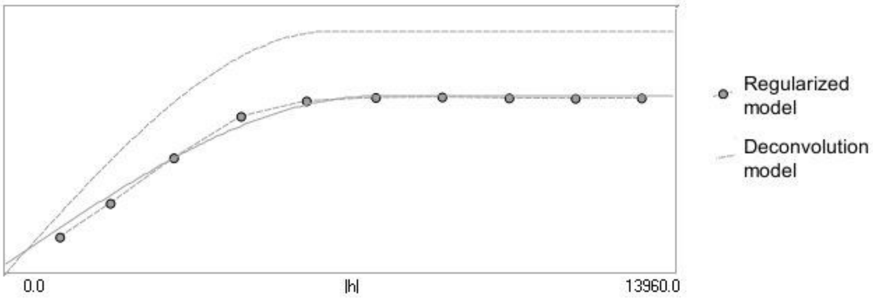


Figure A- 15: Income Deconvolution Model

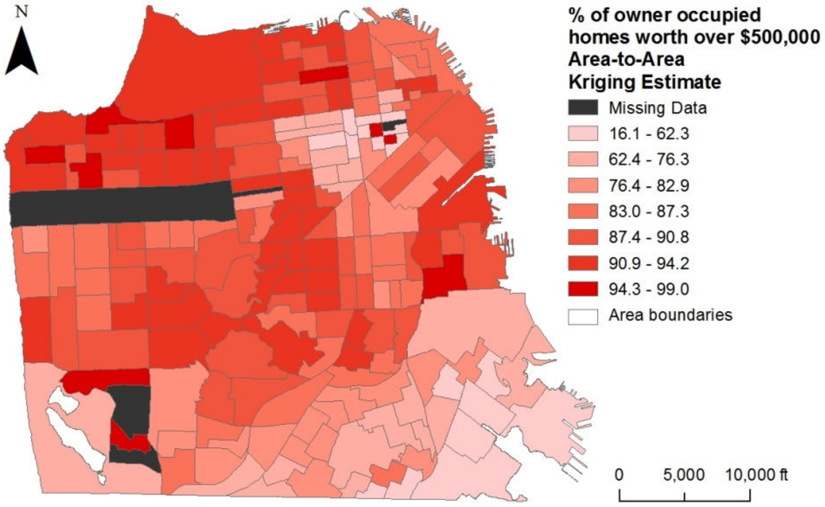


Figure A- 16: Percentage of owner occupied homes worth over 500,000 dollars, Area-to-Area Kriging Estimate

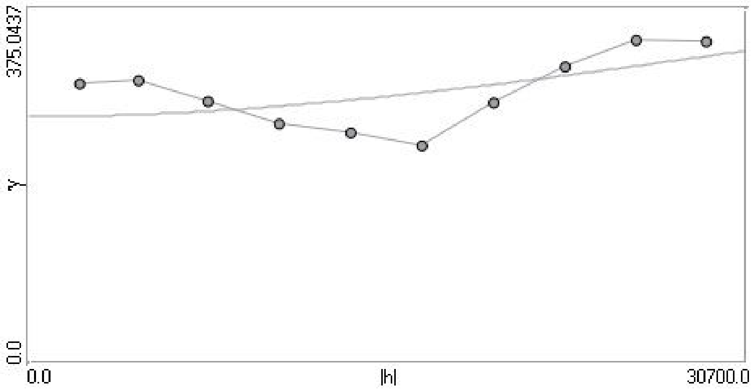


Figure A- 17: House Value Variogram

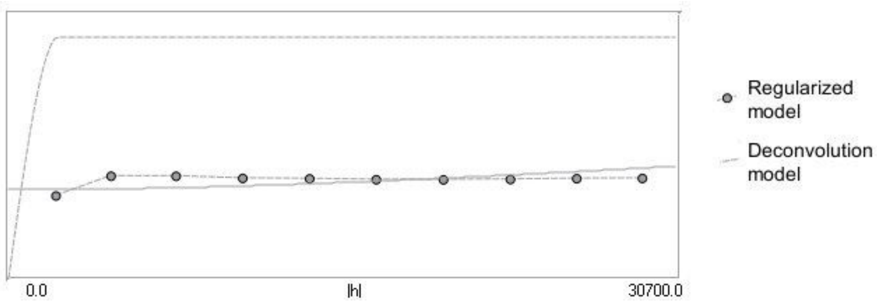


Figure A- 18: House Value Deconvolution Model

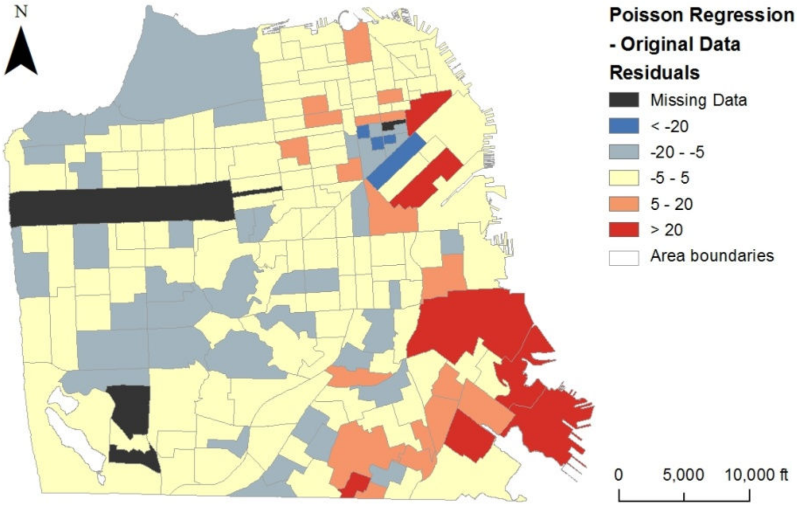


Figure A- 19: Poisson Regression: Original Data – Residuals

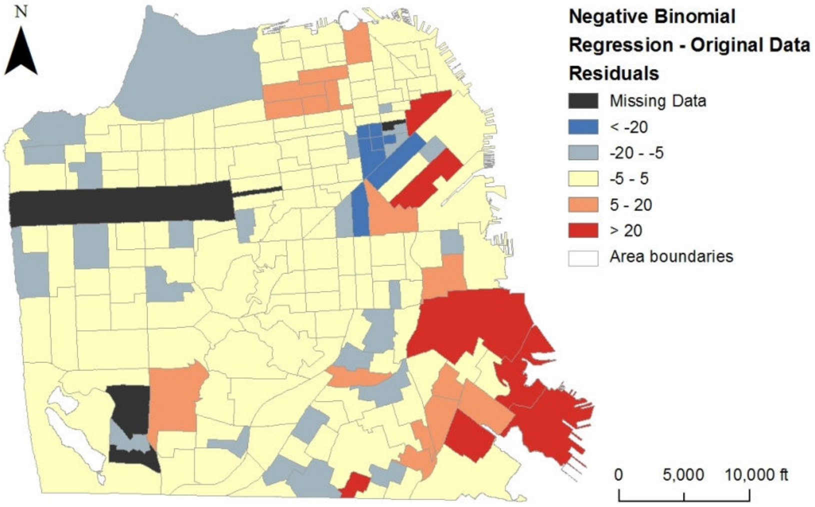


Figure A- 20: Negative Binomial Regression: Original Data – Residuals

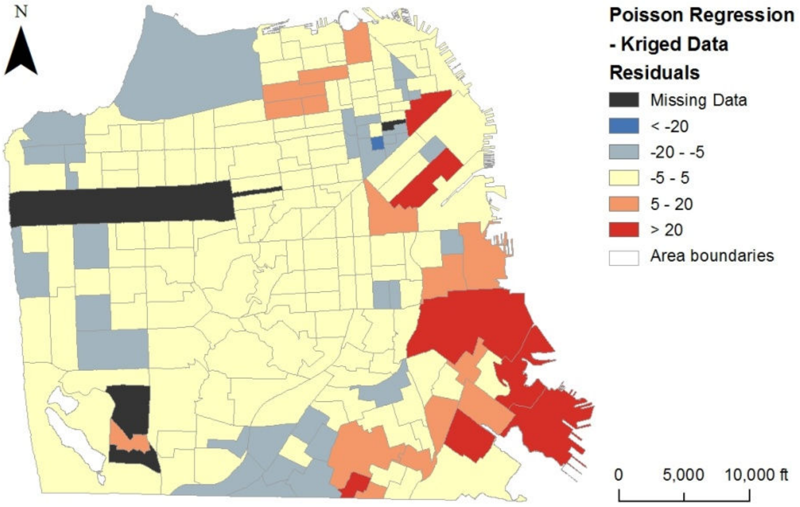


Figure A- 21: Poisson Regression: ATA Kriged Data - Residuals

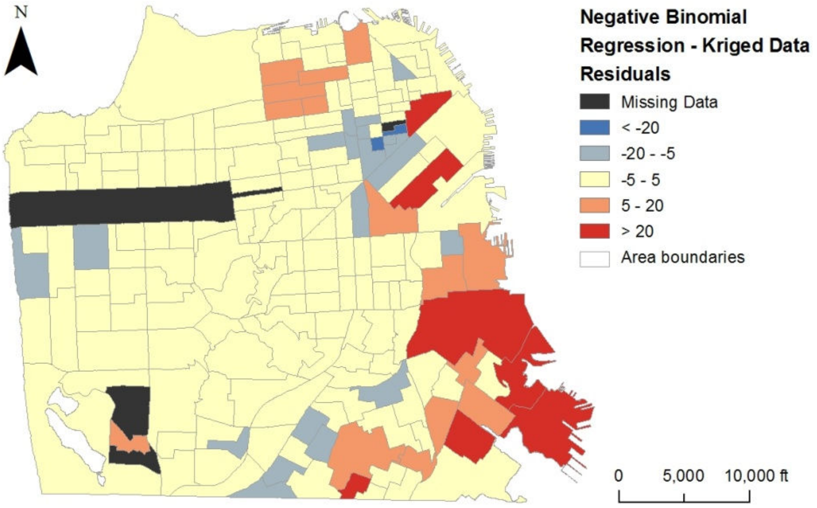


Figure A- 22: Negative Binomial Regression: ATA Kriged Data – Residuals

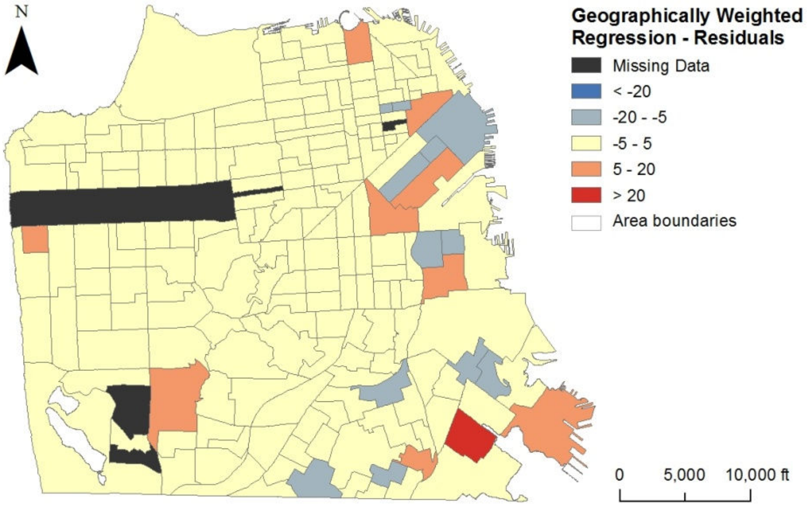


Figure A- 23: Geographically Weighted Regression: Original Data - Residuals

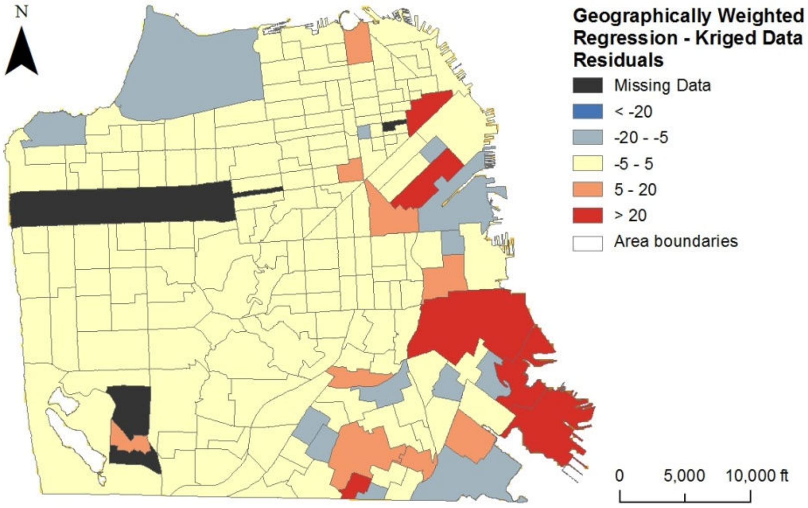


Figure A- 24: Geographically Weighted Regression: ATA Kriged Data - Residuals

APPENDIX B

R code used to carry out the regression analyses:

```
#Open libraries
library(car)
library(pscl)
library(DCluster)

#Set working directory
setwd("~/Thesis/Data/Regression/Working")

#Set data source
Burg <- read.csv("TheVariables.csv", header=TRUE)

#Regression equations
NB <- glm.nb(BurgRate ~ DrugRate + Inc_A + HD_A + Edu_A +
Value_A, data = Burg)
summary(NB)
P <- glm(BurgRate ~ DrugRate + Inc_A + HD_A + Edu_A + Value_A,
family = "poisson", data = Burg)
summary (P)
K_NB <- glm.nb(BurgRateKE ~ DrugKE + IncA_KE + HDA_KE +
EduA_KE + ValueA_KE, data = Burg)
summary(K_NB)
K_P <- glm(BurgRateKE ~ DrugKE + IncA_KE + HDA_KE + EduA_KE +
ValueA_KE, family="poisson", data = Burg)
summary (K_P)

# Variance Inflation Factor (multicollinearity test)
vif(NB)
vif(P)
vif(K_NB)
vif(K_P)

# Pseudo r-squared values
pR2(NB)
pR2(K_NB)

# Overdispersion tests
DeanB(P)
DeanB(K_P)
```