# EVALUATING THE MAUP SCALE EFFECTS ON PROPERTY CRIME IN SAN FRANCISCO, CALIFORNIA

by

Benecia Zahrani

A Thesis Presented to the
FACULTY OF THE USC DORNSIFE COLLEGE OF LETTERS, ARTS AND SCIENCES
UNIVERSITY OF SOUTHERN CALIFORNIA
In Partial Fulfillment of the
Requirements for the Degree
MASTER OF SCIENCE
(GEOGRAPHIC INFORMATION SCIENCE AND TECHNOLOGY)

August 2020

# Dedication

To my family

## Acknowledgments

I am grateful to all the professors who taught me throughout this program, you have helped make a longtime goal a reality. Thank you to Professors Ruddell and Oda, who helped me start and end my thesis journey. Professor Oda provided invaluable support and guidance throughout the process. I would also like to thank my thesis committee members, Professors Wilson and Fleming, whose critiques helped me better formulate why the MAUP matters. Writing is a painstaking process for me, and without the support and patience of my advisor and thesis committee, this thesis would be challenging to read. I appreciate the time and effort they spent proofreading my manuscript.

# Table of Contents

# List of Tables

# List of Figures

# Abbreviations

| | |
|---|---|
| AICc | Corrected Akaike Information Criterion |
| ACS | American Community Survey |
| CA | California |
| CPS | Current Population Survey |
| FBI | Federal Bureau of Investigation |
| GEOID | Geographic Identity Codes |
| GIS | Geographic Information System |
| GISci | Geographic Information Science |
| GLR | Generalized Linear regression |
| GWR | Geographically Weighted Regression |
| LISA | Local Spatial Autocorrelation Method |
| MAUP | Modifiable Aerial Unit Problem |
| MGWR | Multiscale Geographically Weighted Regression |
| OA | Output Area |
| OLS | Ordinary Least Squares |
| SD | Standard Deviation |
| SF | San Francisco |
| SSI | Spatial Sciences Institute |
| USC | University of Southern California |
| MTC | Metropolitan Transportation Commission |

# Abstract

The Modifiable Areal Unit Problem (MAUP) is a phenomenon that occurs when data is arbitrarily aggregated or partitioned to spatial boundaries or units. This phenomenon occurs in most, if not all, spatial analysis efforts. The MAUP effects on analytical results cannot be predicted. The MAUP can cause analysis results, especially statistical results, to vary depending on the scale, aggregation, or partition used for analysis. This fact implies that inferences made based upon the results may not exist if the scale, aggregation unit, or partition change. Yet, in most spatial analysis efforts, there is no consideration of the MAUP. This study explored the MAUP scale effects on the results of optimized hot spot analysis and generalized linear regression analysis by using sociodemographic data and 2017 property crime incidents data in San Francisco, California. A comparative study was conducted at the census block group and census tract scales. The results suggested the presence of the MAUP in the statistical analyses. This thesis provides a framework for geospatial analysts to evaluate the MAUP. It also serves to highlight how the MAUP is present with commonly used analytical methods and data. The results of this research contribute to the body of literature regarding the MAUP effects on cluster and regression analysis.

# Chapter 1 Introduction

This study aims to evaluate the effect of the modifiable unit area problem, known as MAUP, on optimized hot spot analysis and generalized linear regression analysis using San Francisco demographic factors and property crime for 2017 as a case study. The MAUP causes spatial analysis results to vary depending on the scale and aggregation unit used. These differences persist even though the study area, data, and methods are the same. Accurate research is essential. Understanding how or if statistical results vary depending on scale and aggregation methods is something all researchers should be informed.

In regression analysis, relationships between the independent and dependent variables may be present at the census block scale, but not present at the census tract scale because of the MAUP effect. The MAUP can cause relationship strengths, whether they are positive or negative, and correlations between variables to vary at different scales. The MAUP effect can also apply to hot spot analysis. Hot or cold spots may be present at one scale or appear more substantial than the phenomenon they are representing, due to aggregation. The goal is to demonstrate if and how unit and scale affect optimized hot spot analysis and generalized linear regression results statistically and visually. This research also provides a framework for geospatial analysts to evaluate MAUP in a commonly used Geographic Information System (GIS) software, Esri's ArcGIS Pro, Version 2.3.

To explore the scale effects of MAUP, San Francisco was chosen as the case study area. San Francisco was chosen because of its high property crime rate. Per the Federal Bureau of Investigation (FBI), in 2017, San Francisco had the highest rate of property crime in the country (Cassidy and Ravani, 2018). San Francisco is an interesting use case due to its compact shape and urban character. San Francisco is both a city and a county with the same area (Figure 1).


Figure 1 Map of San Francisco County, California. Source: Esri 2020

## 1.1. Why MAUP Matters

The MAUP is a known issue for geographic, health, socioeconomic, and environmental researchers, yet there is no agreed-upon or right way to model or address. While there is no agreed-upon way to research or model MAUP (Montello and Sutton, 2013), researchers must understand that the selection of scale and analysis unit or zone is one of the most integral parts or

their research process (Saib et al., 2014). Due to the MAUP, results can change depending on the scale or zone used for analysis, thus introducing the risk that inferences made based upon those results may not be complete or correct. The MAUP will be present in most spatial analysis and should be considered in all analysis efforts (Swift, Lin, and Uber, 2014). Understanding the impacts of the MAUP on data aggregation at different scales, such as nested census boundaries, is essential. For example, census data is commonly used in regression analysis to help explain rates of crime (Hart and Waller, 2013). Knowing if statistical relationships between explanatory variables, and their contribution to the dependent variable, change as aggregation scales change, contributes to the validity of the research (Flowerdew, 2011).

Understanding the MAUP is critical, especially at the onset of a project when considering what scale and delineation to use, as well as the spatial unit of the data being input into the models. The impacts and intricacies of the MAUP effect on phenomena modeled in statistics and spatial analysis are best described below by Ariba and Petrarca (2011). They suggest the MAUP arises due to aggregation and causes changes to statistical measures. If the scale changes, the statistical measures and inferences made based upon them also change. Not understanding if the MAUP may be an issue in one's spatial analysis and statistics can lead to bias in hypothesis testing and therefore making wrong conclusions.

## 1.2. Geography and Scale and MAUP

Due to the MAUP, it is crucial to understand how relationships between phenomena change as scales or boundaries change. We use geography and maps to understand the world around us. Maps are representations of geographic concepts or events at a specific scale. How we see geographic space or phenomena changes with the scale at which viewed. To view geographic concepts or events at a scale of 1 to 1 is impossible. Geographic concepts, phenomena, or events

are often generalized, combined, or aggregated, as the likelihood of the information being similar or homogenous increases with proximity. This is captured in First Law of Geography, which states, “all things are related, but nearby things are more related than distant things” (Tobler, 1970). Scale, as it relates to Tobler’s law, is important in spatial analysis, because as the scale or aerial unit used for analysis increases, there is less chance that the phenomena are homogenous and the spatial dependence may decrease.

Geographers not only look at geographic space or phenomena. Geographers use the scale as a mechanism to identify patterns and processes that are taking place within that space (Mackaness, 2007). It is crucial to understand how patterns and processes change as the scale used for identification changes. In most cases, understanding geographic patterns or operations is not accomplished at the individual point level. Often point data is aggregated or joined to other spatial structures or boundaries for spatial analysis and statistical modeling. This practice often occurs due to model variables being available within boundaries, such as census data, and a common spatial structure is needed to make assessments. Depending on the boundaries and the scale used to aggregate the data for spatial analysis, the results may differ. When the schemes of aggregations are changed, the analytical results may vary. Changes can occur even if the same data, study area, and spatial analysis methods are used, and are an effect of the MAUP. Phenomena represented by data may have patterns, and cause and effect relationships at one scale that are or are not apparent at another scale due to the MAUP.

The likelihood of finding correlations between data increases as aggregation increases. For example, when evaluating relationships between two variables for a regression model at two nested hierarchal scales, at the smaller aggregation unit, they may not be statistically correlated. However, as the aggregation level or unit increases in size, the likelihood that those two

variables will be correlated increases (Montello & Sutton, 2013). The increase in correlation is an effect of the MAUP due to aggregation to larger enumeration units, and a reason why understanding how the MAUP impacts statistical and analytical results matters.

## 1.3. MAUP Definition and Examples

The two main problems that MAUP has, when data is aggregated or partitioned to boundaries or aerial units, are the scale effect and the zoning effect. These effects occur when the values of aggregated data change depending on how the data is grouped or classified to polygons or partitioned to aerial units (Bolstad, 2016). Factors that play a role in these effects are the shape, size, delineation, and location of the aerial units (Bolstad, 2016). The MAUP is a direct effect of changes in the unit used in spatial analysis. It thus impacts statistical results, highlighting the issue that statistical results are sensitive to boundary changes (Duque, Laniado, and Polo, 2018). Compounding the MAUP issue is that typically spatial data is a representation of some phenomenon taking place that may or may not apply to a whole study area. It is common to use point data to represent geographic events. Then the point data is partitioned into arbitrary spatial units or aggregated to census boundaries for analysis. Often point data is aggregated or partitioned to aerial units that have no relationship to the phenomena it represents (Mennis, 2019).

Figure 2 depicts examples of the scale and zoning effects of MAUP using median age data. In the top left, the figure illustrates the median age by census block and the bottom left by a different aerial unit. In the top right, the data is aggregated to zones that are somewhat equal in size, the same with the bottom left, but the delineation of the zones changed, so they have different median age values. On the top right, the left-most zone has a median age of 39.8, but if the census block data is referred to, the bottom half of the area contains many census blocks with

a median age ranging from 0-30. All of these examples use the same underlying data. None are represented in the same way or show the same results (Bolstad 2016).

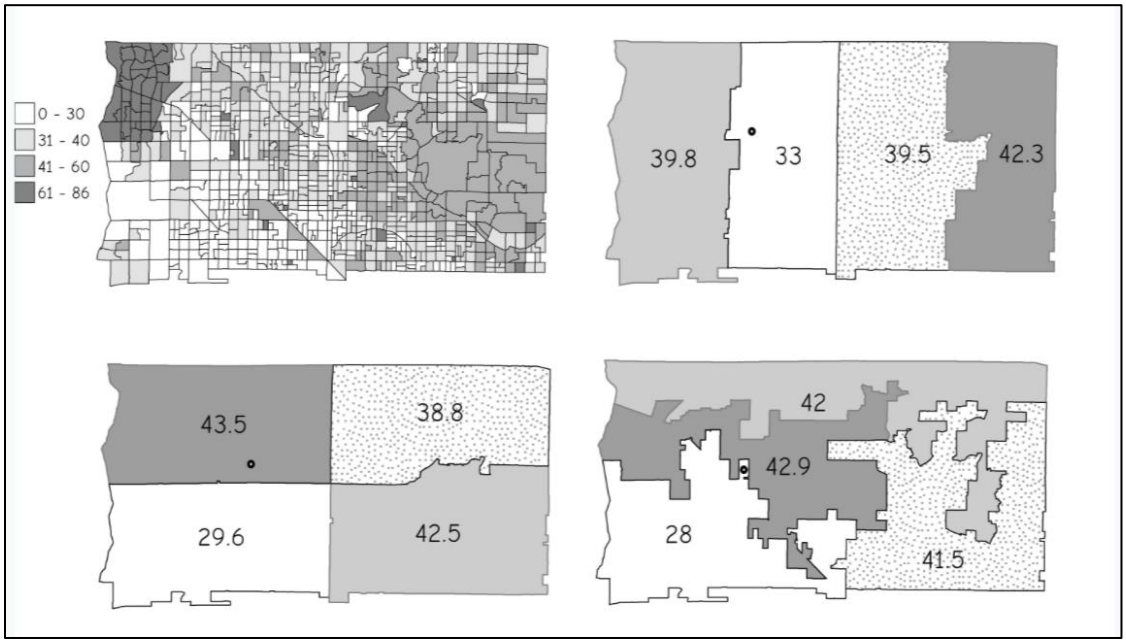


Figure 2 MAUP examples. Source: Bolstad 2016, pg. 392

*1.3.1. Scale and Zoning Effects of MAUP*

Though similar, there are differences between the scale and aggregation effect and zoning effect of MAUP. The scale effect is regarded as a size problem (Lloyd, 2014). Scale effects occur when data is "partitioned" to larger areas than initially captured. A scale effect example is when census data obtained at the household level is represented at the census block level and then aggregated and analyzed at the census tract level (Mennis, 2019). Figure 3 shows the scale effect on the top left and right. The study area is broken into quadrants on the left with select values in each; on the right, each quadrant is broken down into another series of quadrants. This partitioning results in a total of 16 with a complete change in the units compared to the top left representation of the data. The bottom left, and right of Figure 3 shows the zone effect. The delineation and arrangement of the study area changes, while the number of shapes stays the

same. The zone effect of MAUP is, therefore, thought of as a shape problem (Lloyd, 2014). When data within a study area has changes to its boundaries, the information within those boundaries change. Previous studies show that the zone effect in boundary changes can cause a range of -1 to 1 in the correlations of variables in regression analysis (Flowerdew, 2011). Although the study area is the same, and the number of zones or partitions that separate the data is the same, the counts per area or zone change, as illustrated in Figure 3.

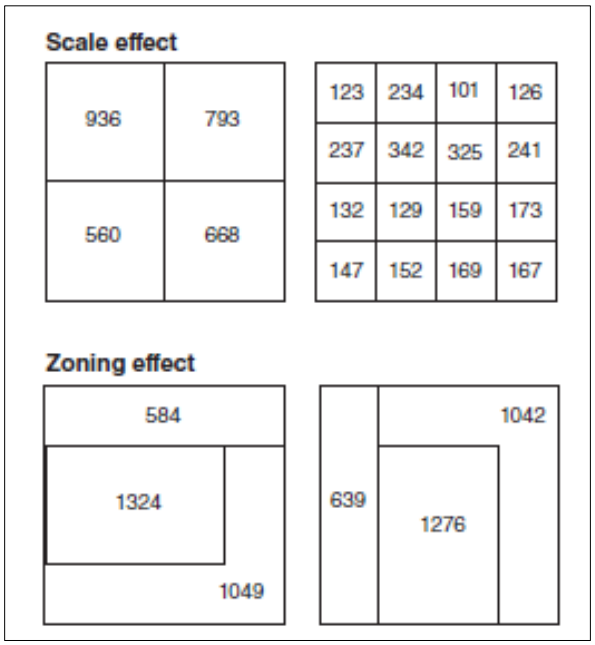


Figure 3 Scale and zoning effects. Source: Lloyd 2014, pg. 30

The population density depicted in Figure 4 at two different areal units highlights the scale effect of the MAUP. On the left, one value represents the population density at the county level scale. On the right, the census tract scale represents population density, and visible differences appear. The area around Birmingham is dense, and the population density declines as one moves farther away from the city center. While the data and study area stay the same, a

change in scale provides a very different representation of population density in Jefferson County.

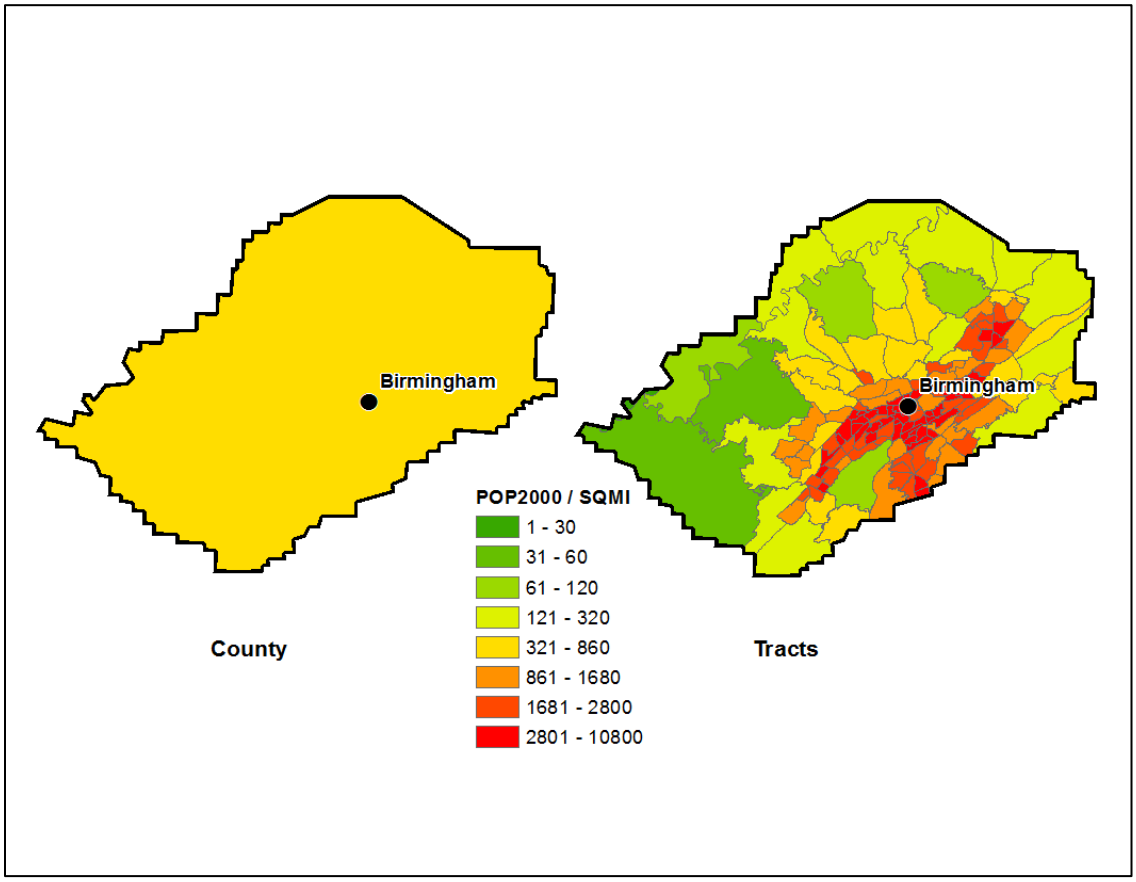


Figure 4 MAUP Scale effect on Jefferson County, Alabama population density

Figure 5 shows an example of the zone effect of MAUP. The same information, household density, is being mapped, but when the delineation of boundaries changes, the result is an entirely different thematic map. The households per square mile range from 91 to 120 using school districts. On the other hand, aggregation based on planning districts shows a higher range from 121 to 150.

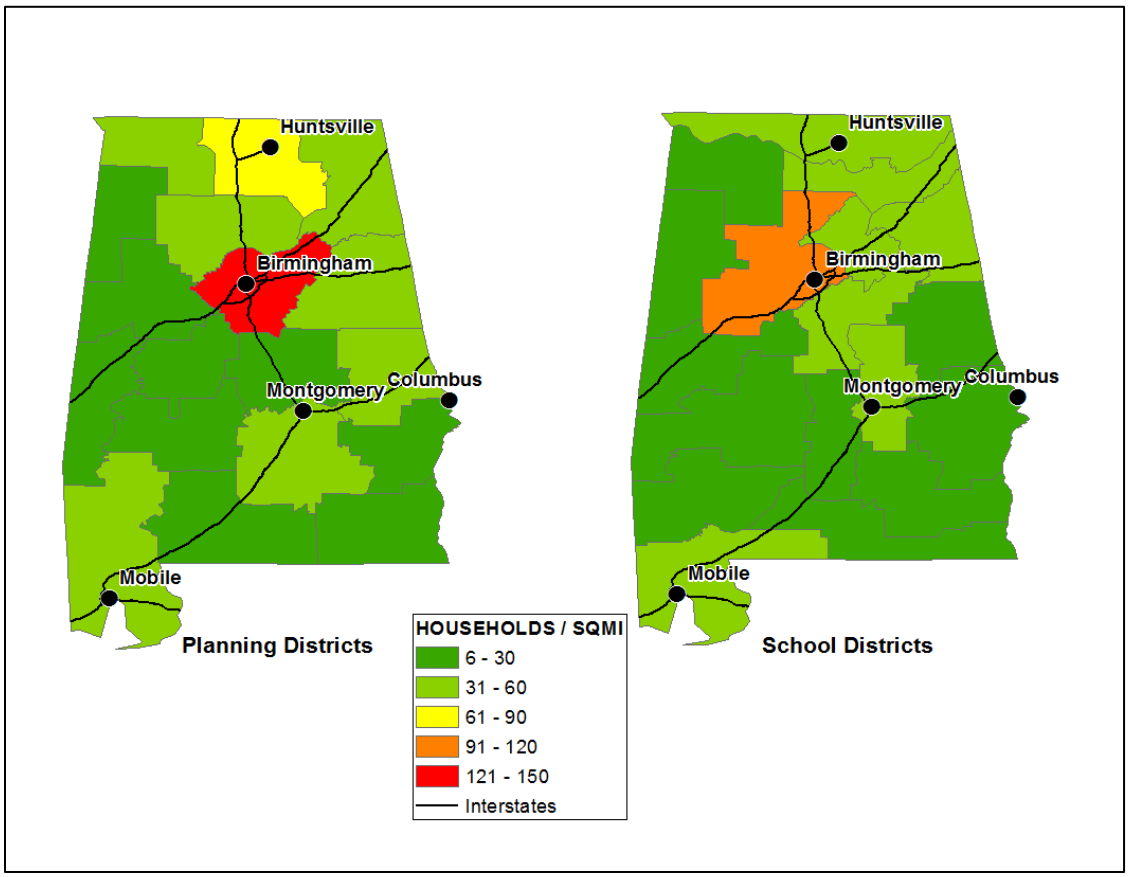


Figure 5 MAUP Zone effect on Jefferson County, Alabama population density

The zone effect results change because of the changes in the delineation of boundaries and how data is aggregated; the scale effect changes results due to the shift in spatial units at each scale. The MAUP causes "variance and covariance of variables" due to the zone and scale effects. An example of the MAUP effect would be seen in a decrease in variance values when aggregating data, resulting in increased correlations. Aggregation results in a "smoothing effect," and if outliers are present, they pull toward the mean (Lee et al., 2016).

### *1.3.2. MAUP's relative – The Ecological Fallacy*

The opposite of the MAUP is the ecological fallacy, which is another type of effect that should be taken into consideration when designing a research project. The ecological fallacy occurs when information or data that applies to a group is attributed to an individual that belongs

to the group (Mennis, 2019). Like the MAUP, the ecological fallacy is a cross-scale inference problem. With the ecological fallacy, inferences about the group or aggregate are applied to conclusions on the individual (Lloyd, 2014). With spatial data, an example of ecological fallacy would be inferring that a census tract value is representative of all the households within that census tract. While the ecological fallacy is not part of this case study, it is a concept to be aware of while conducting spatial analysis and models because statistical results based on inferences from a group may not apply to the individual.

## 1.4. Application of the MAUP to the Case Study

Evaluating demographic data at multiple scales and their statistical relationship to crime can help guide crime interventions. For example, if an indicator of crime is a statistically significant indicator at one scale but not at another scale, how will authorities be able to make inferences about crime, address incidents, or adequately allocate resources to combat crime? The unit of analysis and geography plays a key role in understanding why crime is occurring in a geographic area. Relationships between independent variables that explain crime incidence can change depending on the aggregate units, and these relationships may switch from positive to negative (Porter, 2011).

The MAUP effect in regression analysis can result in changes in the explanatory power of variables and their relationship to the dependent variable. Generally, the effects that the MAUP causes in statistical analysis vary and are often unpredictable. The statistical significance of a variable can depend on the scale of analysis used. Missing essential insights is a possibility if there is no effort to evaluate the effects of scale changes. Researchers should always consider the MAUP when they conduct spatial analysis. The scale of analysis has a direct impact on statistical analysis results. For example, correlations between crime and the variables that constitute

explanatory factors of crime may be different at the census block group and census tract scales. Statistical results also may depend on the zone or configuration that is used for analysis, even if the scale is the same (Flowerdew, 2011). Though both scale and zone are effects of the MAUP, this case study explores the scale effect. The zone effect of the MAUP is not considered in this case study.

The MAUP can also affect hot spot analysis depending on the aerial unit and scale used. For example, if hot spot analysis is performed at the census block group scale, and then repeated at the census tract scale, there can be completely different results. Clusters of hot spots or cold spots can disappear, or estimation of hot spots or cold spot areas can increase or decrease in size, as the area or scale increases.

## 1.5. Thesis Organization

The remainder of this document contains four chapters. The next chapter discusses how the MAUP affects statistical, cluster, and regression analysis. Recent attempts to solve the MAUP are also reviewed. Also discussed in Chapter 2 are crime theories and associated variables that are used to explain property crime. Chapter 3 provides a detailed overview of the data used in this thesis, its preparation for analysis, and the methods which explored the scale effects of the MAUP. Chapter 4 discusses the cluster and regression analysis results. Last, Chapter 5 provides a discussion on the significance of the results and limitations of this study and recommendations for future work.

# Chapter 2 Related Work

This chapter provides an overview of the MAUP and various aspects of crime theory relevant to the methods used in this case study. The literature review suggests the MAUP is an issue that should be considered and explored in every project that deals with geography, scale, and statistical analysis. Implications of why the MAUP is important are seen in how statistical, cluster, and regression analysis results vary depending on the scale or zone used. It is critical to understand how the MAUP effects spatial analysis results.

The discussion is organized into two sections. The first discusses the origins of the MAUP and the MAUP's effect on statistics, regression, and cluster analysis. Recent advancements in attempts to solve the MAUP are also discussed. The second section provides an overview of the crime theories used in this study to guide the selection of regression analysis variables. Crime theories mostly focus on specific categories of crime, and the underlying motivations or socioeconomic processes in an area that can predict its occurrence.

## 2.1. MAUP

The origins of the MAUP date back to Gehlke and Biehl (1934), who found that correlation coefficients in regression analysis increased as the aggregation unit of the data increased when contiguously census tracts were combined. When the census tracts were not contiguously grouped, but randomly arranged the coefficients did not increase in the same manner. The researchers' analysis demonstrated the way data is grouped influences correlation coefficients due to the MAUP. Even though the origins of the MAUP can be traced to 1934, Openshaw and Taylor (1979) coined the term. They determined that regression correlation coefficients of Republican and elderly voters in Iowa varied from 0.98 to -0.81. The range in correlation coefficients changed depending on the scale and enumeration units used to aggregate

data. Their analysis supported the results of Gehlke and Biehl (1934). The MAUP is an acknowledged research problem, especially in the statistics and spatial analysis fields. An extensive body of research, focusing on the MAUPs scale effect on statistical analysis, acknowledges the MAUP's presence and variability in modeling are challenging to predict unless the analysis is conducted at multiple scales. Openshaw and Taylor (1979) laid the foundation for which a growing body of MAUP research is based on by highlighting the MAUP effect on correlation analysis (Wong, 2009). The MAUP is a problem for statistical analysis when data based upon aerial units are used. Thus, the MAUP is present when correlation and regression analysis have a spatial component (Wong, 2009).

The following sections provide a literature review of studies related to the topics and methods addressed in this thesis, the MAUP scale effect on forms of statistical analysis, such as correlation analysis, regression analysis, and cluster analysis. Most research on the MAUP focuses on correlation and regression analysis. Recently cluster analysis has become a topic of interest, as well as advancements in attempts to solve or account for the MAUP.

### *2.1.1. Statistical Analysis Sensitivity to the MAUP*

The MAUP is present in all statistical and spatial analysis research, because of the scale and zone effects. The nature of scale or zone can lead to changes in analytical results and derived patterns (Openshaw, 1977). Statistical analysis results are sensitive regarding the aerial unit in which the data was collected. Due to statistical analysis sensitivity to the MAUP, results using data that has been aggregated to areal units or partitioned to zones may not be reliable (Fotheringham and Wong, 1991). As a result, MAUP has become a topic of interest in health, crime, and economic modeling because these subjects rely heavily on assessing neighborhood effects and processes, both of which can be impacted by the MAUP. Spatial analysis relies

heavily on statistical and mathematical equations. Scale impacts statistics and mathematical equations that are used in spatial analysis (Openshaw, 1977). The impacts often emerge when data is aggregated into higher hierarchal aerial units. The results of statistics depend on the values of the aggregated data and the values around them. For example, when an area that has a low value is surrounded by areas with high values, upon aggregation, the value for the area increases due to the MAUP scale effects. The opposite occurs when an area has a high value surrounded by areas with low values, upon aggregation the scale effects result in decreased values (Wong, 2009). The correlation among variables and their statistical results will strengthen as a result of data aggregation at coarser scales (Mennis, 2019). The MAUP effects on statistical analysis results depend on the data, study area, and scale used for analysis.

Comparing the descriptive statistics of variables at different scales is a way to evaluate the MAUP. A study by Flowerdew (2011) evaluating the MAUP effect on the 2001 Census of England examined the strength of the MAUP between variables at three different scales. The researcher calculated the standard deviations, means, and bivariate correlation coefficients of 18 variables and compared at three sets of census enumeration units. The researcher then classified them based upon mean scale difference, which is the ratio of the mean to the standard deviation. Where there were higher mean scale differences, there was a higher expectation of seeing the MAUP affect the relationship (Flowerdew, 2011). The effects of the MAUP were not as impactful in the study area as predicted. However, the author stated that the MAUP effects seen in the study could be dependent on the spatial autocorrelation of the variables at each scale. Flowerdew (2011), concluded the MAUP could be impactful and yet hard to predict when it will be.

Correlation analysis is sensitive to the MAUP. Changes in correlation results depend on the scale and data format used. Pietrzak (2014) conducted correlation and regression analysis at two scales to examine the MAUP scale effect on the relationships for numerous economic variables. The author explored two dependent variables; total investment outlays in enterprises per capita and the number of entities of the national economy per capita in Poland. The explanatory variables were the number of unemployed, the size of the economically active population, the total investment outlays in enterprises, and the total population. Correlation coefficient increases were seen as the aerial unit increased for data expressed in absolute quantities (i.e., for the variables that were not normalized). On the other hand, correlation coefficients for variables that were normalized per capita data did not lead to changes in the correlation means, but the standard deviation increased significantly (Pietrzak, 2014). Regression analysis was performed on the normalized data. The analysis resulted in large increases and significant changes in regression parameter values and standard error values. The changes of standard error values indicated a high level of variance in the statistical significance of the regression parameters. Based on the results, the authors confirmed the MAUP was introduced to the correlation and regression analysis, when a change of scale was implemented. The authors suggest that analysis using non-normalized and normalized variables should also be considered.

As stated earlier, the MAUP can have effects on statistical results, but those effects can change depending on the data, study area, and unit or partition used for analysis. Swift, Liu, and Uber (2008) conducted a correlation analysis to explore the scale and aggregation effects of the MAUP on relationships between water quality and gastrointestinal (GI) illness. Pearson's r correlation analysis was performed on multiple sets of aerial units: census boundaries, grids, and Voroni tessellations. The correlation values increased from 0.47 to 0.81 as the aerial unit

increased in size. Their findings are similar to the findings of Ghelke and Biehl (1934). The correlation values increasing with the larger aerial units may be due to the smoothing effect of the MAUP, which caused a decrease in the heterogeneity of the data.

### *2.1.2. MAUP Effect on Cluster Analysis*

Another type of statistical analysis the MAUP affects is cluster analysis. The MAUP effect on cluster analysis has recently become a topic of interest for researchers. Clusters are groups of phenomena or data points that are more similar than other groups of events or data in an area. Hot spot clusters are areas where the mean of the data or phenomena modeled is higher than the mean values of other clusters in the area (McKay, 2018). Performing cluster analysis at multiple scales can help with understanding how the MAUP affects cluster analysis.

To evaluate the MAUP scale effect, McKay (2018) used four different clustering methods to identify crime hot spots at two scales: the data zone level and the output area level. The four methods were: k-means, finite mixture models, Local Moran's I, and Getis-Ord Gi*. These cluster methods produced different results, and within each technique, results changed depending on the scale used. For example, the optimized hot spot analysis method using the Getis Ord Gi* statistic produced different results of Strathclyde, Scotland, at each scale using the same crime data (Figure 6). The data zone level deemed the southernmost eastern region as not significant, but when the scale changed to output areas, the same area turned into a crime hot spot, indicated in red. Other areas in Strathclyde at the data zone level contain large cold spots, indicated in blue, but at the output area level, they are not present. The results of this study confirm MAUP presence, and that cluster analysis is sensitive to the scale used for data aggregation. McKay (2018) suggested that the MAUP can cause an incorrect understanding of where crime hot spots persist. Therefore, MAUP can also affect crime mitigation efforts.

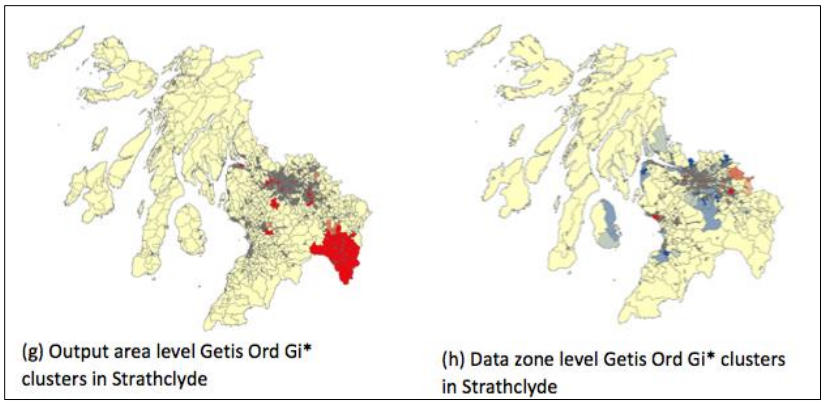


Figure 6 Optimized hot spot analysis of crime in Strathclyde, Scotland. Source: McKay 2018, pg.89

Another study assessing the scale effect of the MAUP on cluster analysis identifies clusters based on industry type. The authors examined the scale effects on cluster analysis by implementing the local spatial autocorrelation method (LISA). Workplace location point data was aggregated to three different nested administrative boundary scales. The workplace types used were advertisement, construction, and stock trading locations. The LISA analysis results were different at each administrative boundary scale for construction workers. At the lowest scale, statistically high valued clusters, significant at the 95% confidence level, were present. At the middle scale, the clusters disappeared, then reappeared again at the next scale up. Nielsen and Hennderdal (2014) confirmed their hypothesis that cluster analysis results of the different business location types would be affected by the MAUP. They acknowledged cluster analysis in the field of economic geography is important for policy and regional development planning, and that the MAUP scale issues in the field are often ignored. They also stressed that the presence of the MAUP should always be considered in cluster analyses.

### *2.1.3. MAUP Effect on Regression Analysis*

One of the first efforts to document the scale and zone effect of the MAUP on multivariate regression analysis was by Fotheringham and Wong (1991). When examining the effect of the MAUP on multivariate regression analysis, they found that the relationship between variables in their model would change depending on the aggregation scale and zone used. Regression models of the Buffalo Metropolitan Area were developed at the census block group and census tract levels. The model hypothesized mean family income would be positively related to homeownership and negatively related to blue-collar workers, the black population, and the elderly. The results of the models at both scales showed differences in parameter strength, significance, and explanatory power. At the census block group level, the black population was significant; however, at the census tract level, it was not. At the census block group level, the $R^2$ value, representing the explanatory power of the variance toward the dependent variable, was 37%. At the census tract level, the $R^2$ value was 81%, more than twice the value at the census block group level. Their discovery was significant because it indicated that multivariate regression analysis results were unpredictable. They stressed that even in a simple multivariate regression model, with a few variables, this would be the case. One of their implications is that researchers should be aware of issues due to the MAUP since they commonly aggregate point data to aerial units, especially in multivariate regression. Fotheringham and Wong (1991) suggested it would be difficult to use the results produced at one scale to inform policy; instead, an analysis should be conducted at multiple scales.

Geographically weighted regression (GWR) has been thought of as a way to lessen the effects of the MAUP because it uses a local regression model instead of a global model. Cheng and Fotheringham (2013), explored educational attainment at two scales, in Northern and the Republic of Ireland through the global ordinary least squares (OLS) regression and local GWR.

The two scales were nested enumeration units at the output area (OA), and ward levels. The GWR model had slightly better results than the OLS model; however, there were still differences in the model parameter estimates, parameter significance, and adjusted $R^2$ values in both models. The OLS model resulted in adjusted $R^2$ values of 0.8 at the OA level and 0.85 at the ward level. In the OLS model, the social class and employment rate variables were statistically significant predictors of educational attainment at the OA level, but employment rates were not at the ward level. The GWR model improved the adjusted $R^2$ values to 0.87 at both the OA and ward scales. The study results imply that GWR for multivariate analysis may mitigate the scale effects of the MAUP compared to the OLS model.

Descriptive statistics and regression analysis were generated by Louvet et al. (2015) in R to illustrate the MAUP scale and data aggregation issues. The study was implemented at multiple scales using normalized and non-normalized variables to explore forest fire incidents. Larger changes in mean-variance and $R^2$ values were seen in models using non-normalized data.

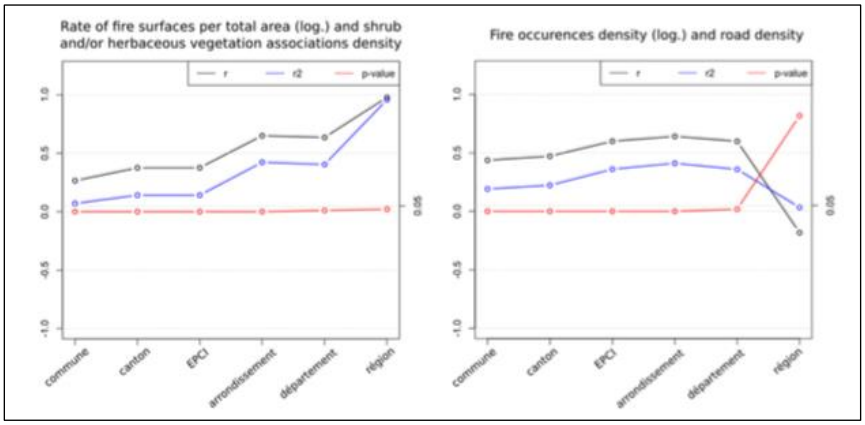


Figure 7 Variation in regression results due to scale. Source: Louvet et al., 2015, pg. 68

$R^2$ values were different for all scales and increased in value as the size of the aerial unit increased. The descriptive and regression analysis results imply that non-normalized variables are more sensitive to changes in scale than normalized variables. Figure 7 depicts the regression

analysis differences at each scale. Louvet et al. (2015) illustrated the results varied at each scale as a result of data aggregation and confirmed the presence of the MAUP.

The MAUP effect on regression analysis may be a result of spatial non-stationarity among multiple predictors and their relationship to a response variable (Parenteau and Sawada, 2011). Spatial non-stationarity is an issue because variable relationships may operate at different scales throughout a study area. The MAUP is a concern for health geography because studies often use census tracts as proxies for neighborhoods, which may not be accurate representations of health processes. The authors state there is a lack of consensus in health geography in terms of which scale is best to model health processes and wondered if the MAUP is a cause. To better understand the MAUP effect on health geography, Parenteau and Sawada (2011) explored the relationship of nitrogen dioxide exposure to respiratory health at three scales. They performed multivariate stepwise regression analysis using 23 variables, selecting the best-fit model for each scale. There was a wide variation in variable coefficient values between the three scales. The number and types of variables that provided the best fit model for each of the three scales were different. The best-fit model for the first scale used six of the 23 variables. The second scale had a best-fit model using four of the 23 variables, and the third scale had a best fit model using seven of the 23 variables. The differences between the best fit models for each scale confirmed the authors' hypothesis that the lack of consensus of the best scale for analysis in the field may be due to the MAUP. The results show how regression analysis of respiratory health issues is sensitive to scale structures and vulnerable to the scale effects of the MAUP.

Saib et al. (2014) provide another health study on the scale effects of MAUP. The authors conducted correlation, local, and global regression analyses at three scales by analyzing the relationship between the mortality of oral and pleural cancer on the one hand and socio-

economic deprivation and environmental exposure via inhalation or ingestion on the other hand. The correlation analysis resulted in adjusted $R^2$ values ranging from 0.24 to 0.28 for pleural cancer mortality and from 0.11 to 0.22 for oral cancer mortality at the three scales. Correlation coefficients for both cancer types and the explanatory variables were different at all three scales. The relationship for two of the variables changed direction at different scales. The regression analysis showed that the local regression model performed better than the global model. Both types of regression models for both types of cancer had differences in the adjusted $R^2$ values, an indicator of model performance, at all the three scales. Saib et al. (2014) attributed the differences in correlation coefficient strength and adjusted $R^2$ values to the MAUP and related data aggregation issues.

### *2.1.4. Attempts to Solve the MAUP*

Some recent research efforts have proposed new ways to test for this. Duque, Laniado, and Polo (2018), for example, have developed the S-maup test. The S-maup test is the first unique statistic created as a way to measure variables and their sensitivity to the MAUP. S-maup is a computation-based method that works by determining the maximum level of aggregation in which a variable can maintain its original characteristics. Fotheringham, Yang, and Kang (2017), on the other hand, have proposed a new version of GWR, called multiscale geographically weighted regression (MGWR), to address computational processes within a study area that may operate at different neighborhoods or scales. MGWR addresses the varying scale of the processes by supporting the use of two or more bandwidths being used in models, instead of choosing one bandwidth, as seen in traditional GWR.

## 2.2. Crime Theory and the Variables

### *2.2.1. Crime Types and Property Crime Definition*

In the U.S., crime incidents are categorized by the FBI Uniform Crime Reporting Program, UCR, as Part I or Part II offenses. Part I offenses are criminal homicide, rape, robbery, aggravated assault, burglary, larceny-theft, motor vehicle theft, arson, and human trafficking. Part II offenses are simple assaults, forgery and counterfeiting, fraud, embezzlement, buying, receiving and possessing stolen property, vandalism, carrying weapons, prostitution, sex offenses not including rape, drug abuse violations, gambling, offenses against family and children, driving under the influence, liquor law violations, drunkenness, disorderly conduct, vagrancy, violating curfew and loitering laws, and suspicion of committing an offense.

Part I offenses are considered serious crimes and are cleared by arrest or other means. They are classified into two categories, a crime against people and crime against property (FBI Uniform Crime Reporting Program, 2017). Property crime offenses are categorized as burglary, larceny-theft, motor vehicle theft, and arson. Property crime occurrence in San Francisco was selected for this case study due to the city having the highest rate in the country in 2017 (Cassidy and Ravani, 2018). San Francisco also had the highest crime rate per capita amongst the largest cities in the United States in 2017 (Cassidy and Ravani, 2018).

### *2.2.2. Independent Variables –Demographic Factors and their Relation to Crime*

Crime incidents typically occur at the neighborhood level and can be tied to demographic and socioeconomic factors in efforts to explain the propensity for crime in an area. There are two main crime theories regarding motivations and structures to explain crime occurrence; social disorganization and routine activity. These theories focus on crime as a construct of social disorganization, poverty, inequality, or a lack of capable guardians to deter crime, thus creating

opportunities to commit crime. Social disorganization theory explores the motivations behind crime, such as poverty, with great differences between the haves and have nots in a study area. Routine activity theory focuses on the premise that there are targets of opportunity in an area.

Routine activity theory is based on the premise that people who live in neighborhoods act as guardians and that they deter crime (Wickes et al. 2016). Routine activity theory focuses on three concepts that are present in space and time for crime to occur: motivation, lack of guardianship, and a suitable target (Moriarty and Williams, 1996). Routine activity theory is based on the premise that crime will occur when offenders believe they will not get caught (Lee and Alshalan, 2005). This would most likely occur when people are away from their homes during the day, or in areas that have suitable targets. Variables commonly used to test routine activity theory are poverty, employment status, presence of multiple housing units, population density, house ownership status, age, and marital status (Lee and Alshalan, 2005).

In contrast to routine activity theory, social disorganization theory is based upon the premise that crime is a result of an unstable neighborhood and an inability to govern. Social disorganization theory assumes crime occurs if the following conditions are present: economic deprivation, residential mobility or population turnover, and racial or ethnic heterogeneity (Cahill and Mulligan, 2007). Variables typically used to test social disorganization theory are poverty, income, number of rental units, percent of single parents, employment rate, population density, ethnic heterogeneity, and education attainment (Andersen, 2006).

Neither social disorganization, nor routine activity theory have been proven alone or together to explain property crime occurrence in an area completely. Moriarty and Williams (1996) and Andresen (2006) stress that it is valuable to look at multiple crime theories when exploring crime occurrence. Andresen (2006) hypothesized crime occurrence in Victoria, British

Colombia, could be best explained using both theories. With this hypothesis in mind, a combination of variables from both social disorganization and routine activity theory was selected for multivariate regression analysis. The variables were ethnic heterogeneity, unemployment rate, population change, number of single parents, average income, population, population density, number of dwellings, young population, percent of college-educated population, and spatial dependence. The regression analysis results incorporated variables from both theories and confirmed the authors' hypothesis that neither routine activity nor social disorganization theory should be used in isolation (Andresen, 2006). The explanatory power of the social disorganization theory was increased by adding the following variables from routine activity theory, the average family income, percent population with a college education, number of dwelling units, and presence of young population age 15-29.

Similarly, Moriarty and Williams (1996) used both crime theories in an attempt to understand property crime victimization. They confirmed their hypothesis that crime occurrence would be higher in socially disorganized areas than organized areas. In addition to this, they also suggested that property crime correlation analysis using routine activity theory would have better results in areas that were more socially disorganized. The correlation analysis in this work used house value, a security index, age, race, homeownership, residential stability, home during the day, home on the weekends, employment status, neighborhood help, marital status, and the number of adults in the home as explanatory variables. As a consequence of the supporting literature, variables commonly used in both theories are selected for the regression analysis part of this thesis study and are discussed next in Chapter 3.

# Chapter 3 Methods and Data

This chapter provides an overview of the methods and data used to analyze how the scale and aggregation effects of the MAUP impact statistical analysis results. Based on the objective and literature review, two spatial analysis methods (optimized hot spot analysis and generalized linear regression) were selected to measure statistical analysis sensitivity to the MAUP on property crime incidence in San Francisco, California. This chapter is divided into two sections. The first describes the process taken to acquire, assess, and prepare the data for analysis. The second describes optimized hot spot analysis, exploratory regression, and generalized linear regression, and how they were implemented. Exploratory regression was used to explore the relationships between the variables and finalize the variable selection for the generalized linear regression.

## 3.1. Data Acquisition, Assessment and Preparation

Spatial analysis was conducted to demonstrate the effect of the MAUP on property crime incidence at the census tract and census block group levels. The impact of the MAUP was evaluated by performing optimized hot spot analysis and generalized linear regression analysis. For these analyses, crime data, boundary data, and demographic data (Table 1) were first acquired, then assessed and prepared for spatial analysis. The datasets used in the optimized hot spot analysis were 2017 property crime incident data and the census tract and census block group boundaries in San Francisco. For the regression analysis, the dependent variable was the 2017 property crime incident data, and the independent variables were selected from 2018 demographic indicators processed by Esri.

Table 1 Datasets and data sources

| Dataset | Spatial Resolution | Temporal Resolution | Data Format | Data Source |
|---|---|---|---|---|
| Crime Data | Latitude / Longitude | 2017 | GeoJSON, point | San Francisco Government |
| Demographic Data | Census Tract & Block Group | 2018 | Geodatabase, polygon | Esri |
| San Francisco Bay Region Census Tracts & Block Groups | Census Tract & Block Group | 2018 | Shapefile, polygon | San Francisco Metropolitan Transportation Commission |

*3.1.1. Study Area and Boundary Data*

The TIGER/Line census boundaries are data created by the U.S. Census Bureau. They represent hierarchal geographic entities, ranging from nation to region, state, county, census tract, census block group, and census block. They do not contain demographic data; instead, they have geographic identity codes (GEOID) that census data or other data sources can be linked with. Census boundaries often cover areas that are not lived in, such as parks and large bodies of water. The inclusion of such areas may lead to erroneous results in analysis. The unclipped TIGER/Line boundaries covered the large water bodies to the east and west of San Francisco (Figure 8). If the unclipped boundaries were directly used for analysis in this thesis, there would be large areas of water with no crime events or population. The city of San Francisco uses a set of 2018 Census TIGER/Line boundaries that have been clipped to remove water. These boundaries were used for the optimized hot spot analysis and regression analysis in this study. San Francisco's clipped boundaries were joined to the Esri demographic data using an attribute join, based on the GEOID unique identifiers of the census tracts and census block groups.

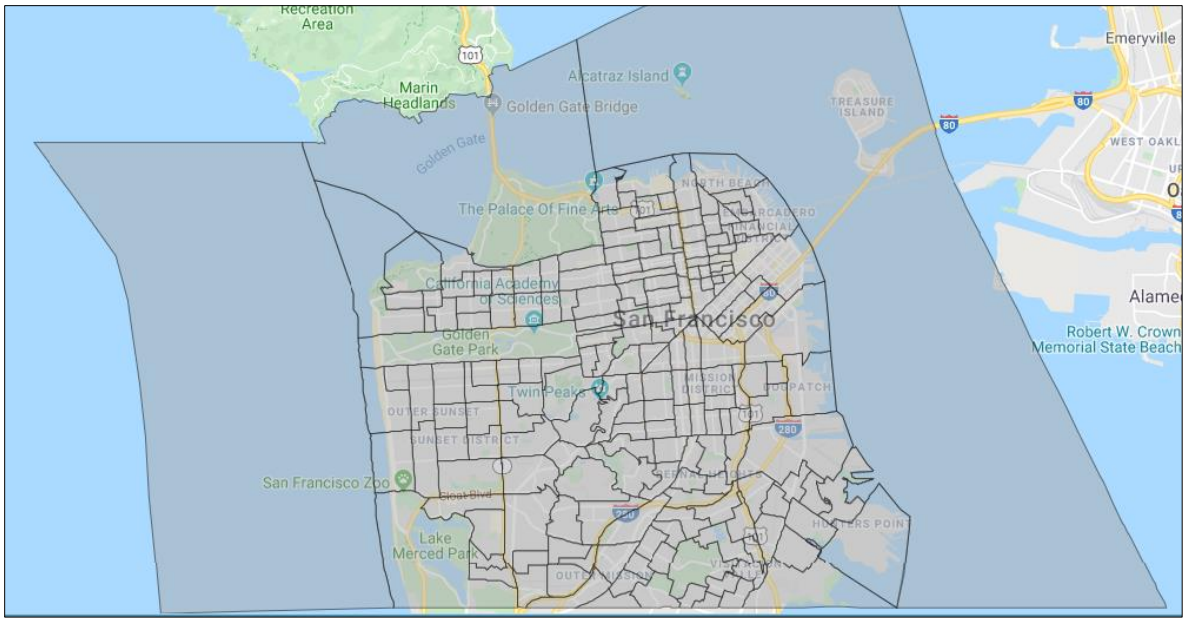


Figure 8 San Francisco unclipped TIGER/Line census tract boundaries

It is also important to note that this study did not include spatial outliers in San Francisco. Golden Gate Park, Census Tract 9803, was omitted from the study because it has low population density and a large area. There is an island group called the Farallon Islands that are part of Census Tract 9804.01, with no population approximately 35 miles west of the conterminous San Francisco. No crime incidents or population have been assigned to this tract. There is another census tract, Census Tract 179.02, that has areas not contiguous to the mainland that has were removed from the study area as well. These areas are Treasure Island, a part of Angel Island, and a parcel of land across the bay touching Oakland. Of the regions in Census Tract 179.02, Treasure Island was the only area containing property crime, with a count of 82 events. The crime events for Treasure Island were deleted. Northern San Francisco also has an area that is not within the San Francisco Police Department jurisdiction, called Presidio, Census Tract 601. Presidio is considered a separate entity and within the jurisdiction of the California State Park Police, because it is federal land. Presidio is not included in the study area because the crime data

were unavailable. All other parks in San Francisco are within the jurisdiction of the San Francisco Police department and were included in the study area, as property crimes occurred within their boundaries.

The resulting study area contained 59,650 property crime incidents, 192 census tracts, and 576 census block groups. All removed census tracts shared the same size and geographic area with the subordinate census block groups. Also removed were the coincident block groups. In the remaining study area (Figure 9), there were eight census tracts and eight census block groups that are entirely coincident. These enumeration areas were kept in this study since most of them contained parks and high population counts.

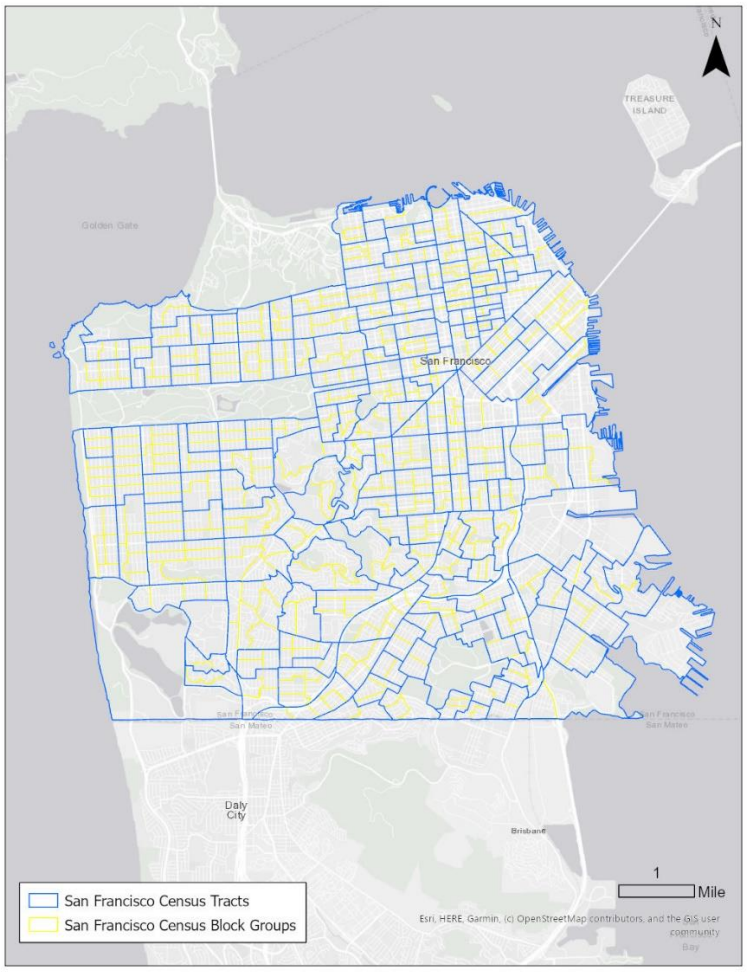


Figure 9 Final study area

*3.1.2. Crime Data*

Initially, the intent was to use the 2018 crime data in this research project. Two subsets of crime data were downloaded from the San Francisco Government open data website in GeoJSON format because the data was not available as a single dataset for the year 2018. The first subset was from the San Francisco legacy crime database, which ranged from January 1st, 2003 to May 1st, 2018. The second subset of data was an updated crime database starting May 1st, 2018. As of May, the city of San Francisco changed their crime database schema and started masking crime incidents to intersections to handle privacy concerns. Instead of the crime point being at the location where the crime incident occurred, it is at the nearest street intersection. Due to the location modification of the crime incidents resulting in degraded spatial granularity, the updated database was deemed not usable for analysis. There would be no way to know which census block group or census tract the incidents should be joined to, as the intersections coincide with the census boundaries. Due to the masking of crime incident locations starting in May of 2018, crime incidents from 2017 were used in this study. Figure 10 depicts the masked crime incidents, with the degraded spatial granularity aligned to street intersections.

For the year 2017, there were a total of 154,773 crime incidents. Before the crime data was appended to the census block groups and census tracts for analysis, it was filtered to contain only property crime incidents. Burglary, larceny-theft, motor vehicle theft, and arson were then selected, resulting in a total of 59,650 property crime incidents in the study area. These incidents were then spatially joined to census tracts and census block groups for the optimized hot spot analysis and regression analysis. The spatial join operation ensured that each enumeration unit contained the sum of crime incidents.


Figure 10 Spatial location of 2018 San Francisco masked crime incidents

*3.1.3. Demographic Data*

The demographic data in this study came from the Esri curated and proprietary Popular Demographics dataset. This dataset was part of the Living Atlas of the World, a repository of geographic data hosted on ArcGIS Online at arcgis.com. The demographic dataset had over 50 unique variables, which were recorded at the census block group up to the state level. The variables include population, household size, employment rate, race, poverty, income, housing unit, housing status, and housing values. The San Francisco data was downloaded from Esri

ArcGIS Online. Close inspection of the demographic data showed that the polygon boundaries did not correctly nest within each other as census block groups and census tracts should. The block groups contained polygons in which the verticies and lines overshot the adjacent census tract boundaries, so the data was not topologically coincident with the corresponding tract. This type of error would cause issues when aggregating crime counts. Due to the nesting issue, the Esri boundaries were not used, and the demographic data variables were combined, with the clipped San Francisco 2018 Census TIGER/Line boundaries discussed in Section 3.1.1.

During the data collection phase of this project, the American Community Survey (ACS) data was considered as a possible source for the demographic variables as well. The Esri demographic data was selected because it had updated and refined 2018 estimates available at the census tract and census block group scales. Esri implements a robust method using multiple sources to create demographic data more accurately to capture yearly changes to the population and geography. The robust method Esri employs is called the Address Based Allocation (ABA) methodology. In an independent evaluation, a panel evaluated the data via comparisons with different four datasets developed by other vendors. Esri's data had the lowest precision errors for the population and household variables. The panel acknowledged that population and household variables are more difficult to estimate for smaller geographies like census block groups (Esri, 2012).

The Esri demographic data can be divided into population and housing characteristics. Multiple data sources such as the 2010 Census, ACS, and Current Population Survey (CPS) and cohort survival models are used to calculate the estimates for the population characteristics. Housing data information is based on the 2010 Census, which was updated by Esri using sources construction data from Metrostudy, Axiometrics, county building and home permits, US Postal

Service (USPS), ACS, CPS, and the Housing Vacancy Survey sources. Changes in home values were tracked using House Price Index (HPI) and the Federal Housing Finance (FHFA) information from mortgage loans provided by Fannie Mae or Freddie Mac. Labor Force and Household income information were derived from the ACS, CPS, Local Area Unemployment Statistics (LAUS), Occupational Employment Statistics (OES), Bureau of Labor Statistics (BLS), and Bureau of Economic Analysis (BEA). Household incomes were updated by accounting for the change in the working population.

### *3.1.4. Data Normalization*

It is common practice to normalize data from counts to ratios before use in statistical analysis. Normalizing data is an approach to limit the magnitude of counts by converting them into rates, which are a measure of intensity (Dailey, 2006). For example, if a city-level crime analysis is performed for a whole state, cities with larger populations are likely to have more crime incidents. If the crime incidents per city are converted to ratios, it minimizes the differences between the smaller and larger cities. Data can be normalized in two ways, by the sum of raw total values or by another associated attribute. For example, the labor force can be normalized by the population over age 16. All demographic variables were normalized by the sum of the total count, resulting in a ratio in this study, except the unemployment rate, diversity index, and median household income variables. The property crime data were also normalized by taking the number of property crime events per tract or block group, then dividing it by the total number of crime events in the study area. Below is a list of all the variables used for the regression analysis (Table 2).

Table 2 Variables used for regression analysis

| Variables | Type |
|---|---|
| 2017 Property Crime | count |
| 2018 Female Population | count |
| 2018 Male Population | count |
| 2018 Unemployment Rate | ratio |
| 2018 Median Household Income | ratio |
| 2018 Median Home Value | ratio |
| 2018 Diversity Index | rank |
| 2018 Owner-Occupied Housing Units | count |
| 2018 Renter Occupied Housing Units | count |
| 2018 Vacant Housing Units | count |
| 2018 Hispanic Population | count |
| 2018 White Non-Hispanic Population | count |
| 2018 Black/African American Non-Hispanic Population | count |
| 2018 Asian Non-Hispanic Population | count |
| 2018 Pacific Islander Non-Hispanic Population | count |
| 2018 Other Races Non-Hispanic Population | count |
| 2018 Multiple Races Non-Hispanic Population | count |
| 2018 Household Income $200,000 or greater | count |
| 2018 Household Income $15,000 or less | count |

## 3.2. Methods Workflow

ArcGIS Pro was utilized for data preparation and analysis in this case study. Once the study area was determined, the data were collected, formatted, and normalized for use as inputs in the analysis (Figure 11). All data was projected to NAD 1983 2011 San Francisco CS13 (ftUS). Optimized hot spot analysis and generalized linear regression analysis were performed at the census block group and census tract scales to answer the research question. Exploratory regression analysis was also conducted to explore the selection of variables for the generalized linear regression. The dependent variable in both forms of statistical analysis was the property crime locations in San Francisco. The demographic variables were used as the explanatory variables for explaining the occurrence of property crime.

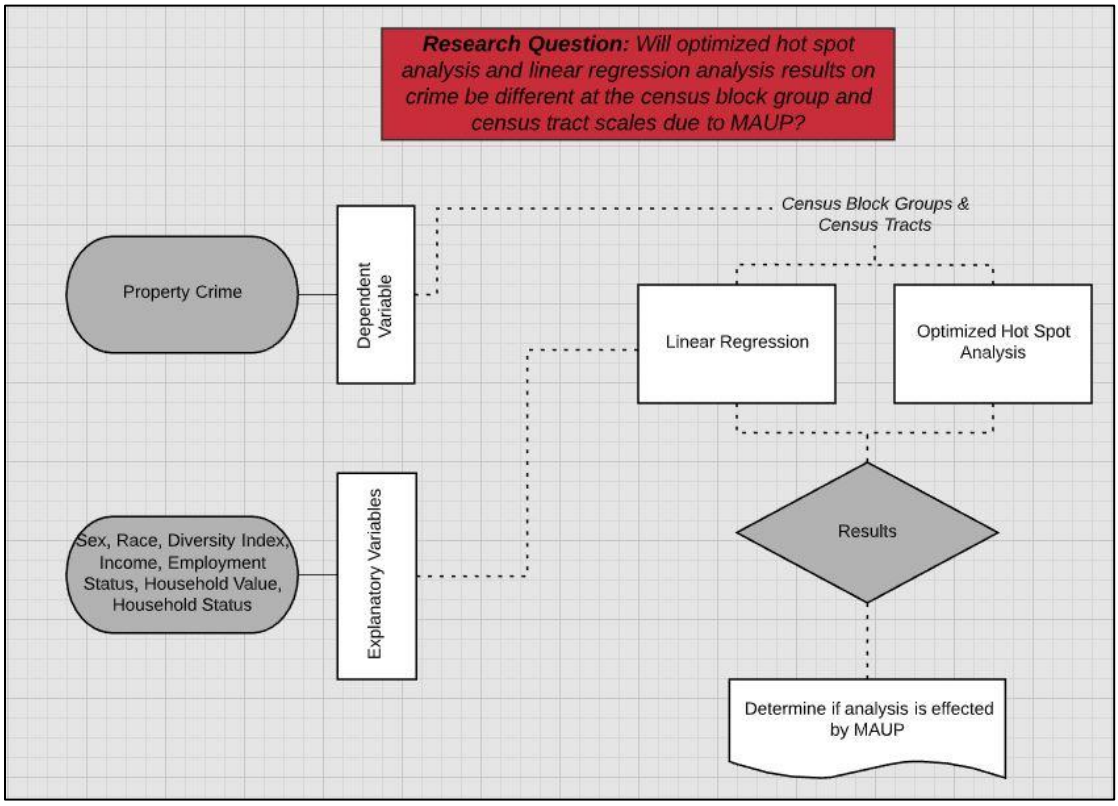


Figure 11 Research question and methods workflow

### *3.2.1. Optimized Hot Spot Analysis*

Optimized hot spot analysis was demonstrated by using the property crime point data to identify hot spots and cold spots at each aggregation boundary level. The optimized hot spot analysis tool provided three options for its areal aggregation: (1) count incidents within the fishnet grid; (2) count incidents within the hexagon grid; and (3) count incidents within the aggregation polygons. The method used in this study was counting events within aggregation polygons. The property crime points were input as the incidents, and the census block groups and census tracts were input as the aggregation polygons. The number of property crime incidents was counted in each polygon, and then, the sum was used by the tool for analysis.

The optimized hot spot analysis tool uses the Getis-Ord Gi* stastistic to determine where phenomena of interest are significantly clustered. Features that have high Gi* values are significant clusters, and features that have Gi* values close to 0 are not significant (Mitchell, 2005). Mitchell (2005) provides the Getis-Ord Gi* as the following equation:

$$G_i * (d) = \frac{\sum_J w_{ij}(d)x_J}{\sum_J x_J}$$

Where $G_i^*$ for a feature $(i)$, at a distance $(d)$ and the value of each neighbor $(x)$, is multiplied by the weight for the target-neighbor pair $(W_{ij})$, and the results summed. Then the sum is divided by the sum of the values of all neighbors $(X_j)$, that is, all features in the data set.

For a location to be considered statistically significant, it has to meet the requirement of not only having a high value, but it also has to be surrounded by other high values, or have a low value surrounded by other low values (Esri, 2019a). Another condition is that the local sum for each feature and its neighbors has to be proportionally higher than the sum of all features in the study area. The optimized hot spot analysis tool automatically determines the optimal scale of analysis to yield the best results. This determination is known as the distance band threshold. For the optimal scale of analysis, the tool uses incremental spatial autocorrelation by computing the intensity of clustering at each feature distance through the Global Moran's I statistic. This process results in a peak distance for the scale of analysis. If no peak distance is found, the optimized hot spot analysis tool examines the distribution of each feature in relation to its neighbors by computing the average nearest neighbor distance, to find the distance band threshold. After the distance band threshold is determined, the Getis-Ord Gi* statistic method is run to determine the statistically significant features.

The outputs of the optimized hot spot analysis tool are z-scores, p-values, and Gi-Bin confidence levels. Gi-Bin confidence levels identify if there is significant clustering of high values – hot spots, or low values – cold spots. Z-scores are provided as standard deviations and are used by the tool to identify if the pattern seen is random or statistically significant. P-values tell us whether the probability of the observed spatial pattern is random or not. A small p-values with either of very high or very low z-score means that the pattern is not random, and indicates a hot spot or cold spot is present (Esri, 2019a).

### *3.2.2. Regression Analysis*

Regression analysis is commonly used to determine the relationships connecting one or more independent variables and a dependent variable. Regression statistically assesses the strength of relationships in the social sciences (Cheng & Fotheringham, 2013). Understanding how the MAUP changes relationships as the scale and aggregation unit changes, will help researchers to be more informed about their analysis and also help stakeholders with decision making. Most regression analyses assume the data are normally distributed. The distribution of all the variables was checked using a histogram in this study. Most of the variables were not normally distributed with positive skews and high kurtosis values. All the variables except median household income, the unemployment rate, and the diversity index were transformed prior to the regression analysis itself using the log function to resolve the skewness. Transforming data is a process in which all the data values are converted to a new scale, thus changing the distribution (Mitchell, 2009).

In this study, two types of regression analyses were conducted. Exploratory regression was performed first to explore the relationships among the variables to select the variables for input into the second regression model. The second step used generalized linear regression which

to create the final model and evaluate the scale effects of the MAUP. Exploratory regression and generalized linear regression use a common regression technique called ordinary least squares (OLS). OLS is a form of linear regression that generates prediction values based on the observed values of dependent variables in relation to explanatory variables (Figure 12). OLS is one of the most common forms of regression and often thought of as a starting point in spatial regression analysis (Esri, 2019b).

OLS uses the following mathematical equation to show relationships between the dependent variable, what is being predicted, and the independent variables:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \varepsilon$$

The ordinary least squares regression equation contains the dependent variable $y$, property crime, which is a function of the regression coefficients, β, for every explanatory variable, $x$, and represents the strength and type of the variable relationship to $y$. The regression intercept, $\beta_0$ is the expected value of the dependent variable if the independent variable is 0. The residuals, $\varepsilon$ , represent the difference between the observed and predicted values in the model and the unexplained portion of the dependent variable (Esri 2019b).

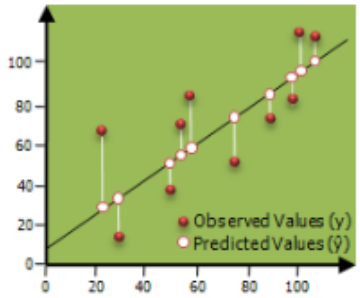


Figure 12 Ordinary least squares: Predicted values in relation to observed values. Source: Esri 2019b

OLS determines the relationship between the dependent variable and independent variables by calculating regression coefficients for every variable. The regression coefficients

indicate the strength and type of relationship between the dependent and explanatory variables (Esri, 2019c.). In this study, the independent variables, demographic factors, are used to predict property crime occurrence, the dependent variable.

#### 3.2.2.1. Exploratory and Generalized Linear Regression Workflows

As stated in the previous section, exploratory regression was used to explore the relationships between the model variables and to determine what independent variables were positively or negatively correlated to property crime at the census block group and census track scales. It resulted in a table with the variable significance rated on a scale from 0 to 100 and multicollinearity values. The exploratory regression results were compared to evaluate differences at each scale. The exploratory regression helped solve multicollinearity issues and narrow down the selection of variables based on crime theories for the final generalized linear regression model.

Generalized linear regression was run next, and the results, correlation coefficients, $R^2$, probabilities, and AIC values were used to evaluate the scale effects of the MAUP. The correlation coefficients represent the relationship type and strength between the dependent and explanatory variables. The $R^2$ values suggest how the model explains the observed dependent variable. An $R^2$ value of 1 means that the model explains 100% of the variation in the dependent variable. An $R^2$ of 0.50 means that 50% of the variability of the dependent variable is explained. The probabilities identify if a variable is statistically significant. AIC values tell how well the model fits the data, the smaller the value, the better. The results of the regression models were evaluated and compared to determine the MAUP effects at different scales.

# Chapter 4 Results

The results presented in this chapter explore the scale effects of the MAUP on the results from two forms of statistical analysis; optimized hot spot analysis, and regression analysis. The applications used two aggregation scales and property crime incidence in San Francisco, California, as the case study. The results support findings discussed in the literature review and confirm the study hypothesis that the aggregation of data to different scales will affect statistical analysis results. This chapter reviews the results of these analyses in detail.

## 4.1. Optimized Hot Spot Analysis Results

The Optimized hot spot analysis results were different at each scale; therefore, the assumption that there would be differences in the results at each scale due to the MAUP scale effect was confirmed. Aggregation to aerial units at different scales produced MAUP scale effects, even when the same data and study area were used.

The optimized hot spot analysis at the census block group scale, resulted in a larger hot spot than at the census tract scale. The optimized hot spot analysis results are provided as maps (Figures 13 and 14) that depict the statistically significant hot spots and cold spots, at three confidence levels, and areas that are not statistically significant in off white. Areas in red signify that there are clusters with high levels of property crime present. The areas in blue indicate that there are clusters with low levels of property crime present. To be considered a hot or cold spot, the area has to have a high or low value surrounded by similar values. When compared to the sum of all features in the study area, the areas that have local sums higher relative to the sum of all the features are considered statistically significant.

There was also a cold spot at the census block group scale. This cold spot is located south of Golden Gate Park, in an area identified as the Sunset District (Figure 13). The cold spot is

surrounded by areas that are not significant. The census block group has a large contiguous hot spot in northeast San Francisco. The hot spot is larger at the census block group scale extending into the Chinatown neighborhood, whereas this area was not significant at the census tract scale. The census block groups had approximately 2.5 square miles more territory that was deemed a hot or cold spot than the census tracts. Not only did the change of scale from block groups to census tracks decrease the size of the hot spots, but it also changed the confidence levels (Figure 14). At the census block group scale, the blocks on the western and northern periphery were statistically significant at the 90 and 95% confidence levels.  In contrast, the hotspot at the census tract scale, displayed a similar confidence level only on the western periphery.

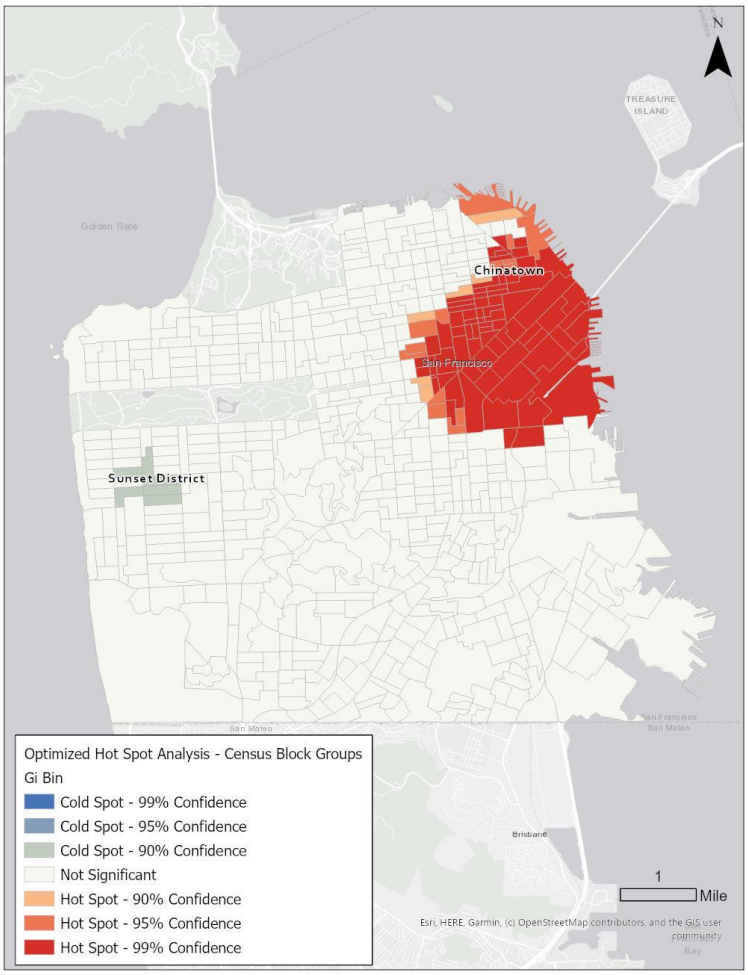


Figure 13 Property crime optimized hot spot analysis at the census block group scale

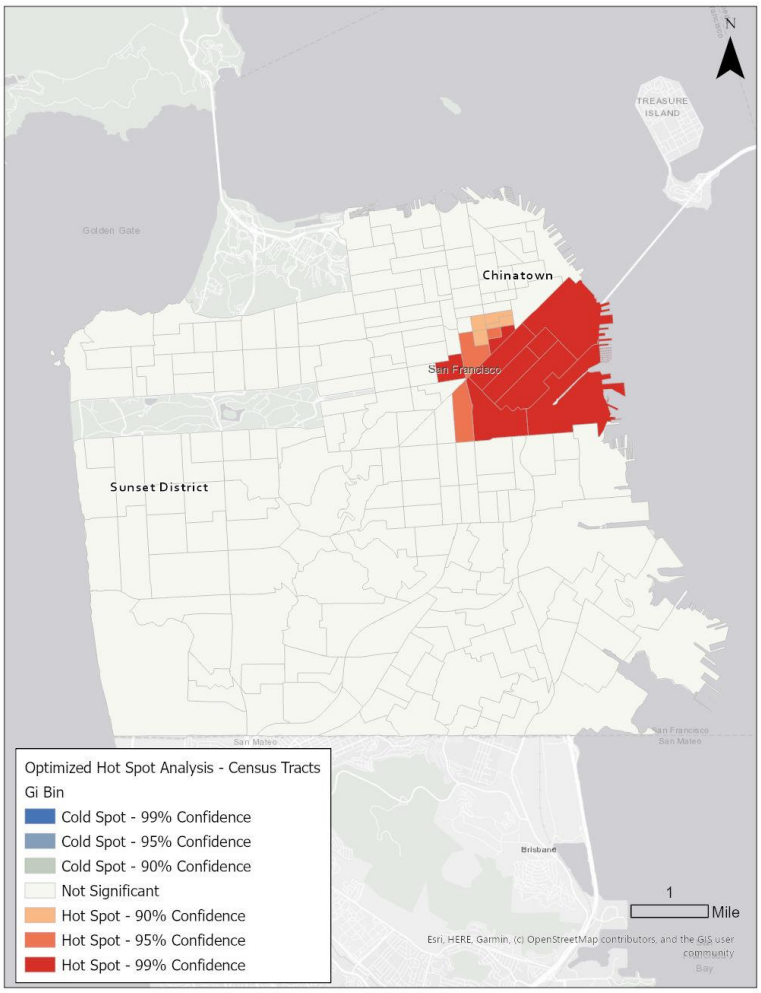


Figure 14 Property crime optimized hot spot analysis at the census tract scale

One reason the optimized hot spot analysis results were different at each scale is because the Getis Ord Gi* statistic used distance band thresholds and feature weights to identify where there were statistically significant clusters in the study area. The distance band threshold is used to determine how many neighbors each feature has. Features that have higher weights than nearby neighborhood features are considered significant clusters. The optimal distance band for the block groups was 4,283 feet, and for the tracts was 4,297 feet (Appendix A). As the scale changed, the distance band threshold changed, and this sometimes caused a change in the number of neighbors for each feature.

## 4.2. Regression Analyses Results

Exploratory regression analysis was used to explore the relationships between the demographic variables and property crime and to help determine what variables to keep for the generalized linear regression.

### *4.2.1. Exploratory Regression Results*

The results produced using the census block groups and tracts were different as expected, confirming that regression analysis is affected by the MAUP. The relationships of some variables to property crime changed, from the census block group to the census tract scale. Exploratory regression was used to test for multicollinearity between all the variables by examining the VIF values. Overall, multicollinearity increased as the aggregation scale increased from census block groups to census block tracts. High VIF values, typically above a threshold of 7.5, mean that there are multicollinearity issues between variables. Multicollinearity issues indicate that one or more of the variables is redundant. Removing these variables will help to resolve the issue. In this study, the Hispanic population variable caused multicollinearity issues with the race and sex variables, resulting in high VIF values. Therefore, all the race variables were removed. Since removing only one race variable did not make sense, the other race variables were removed as well. Instead, race variations were accounted for in the Diversity Index variable.

Some census tracts had a value of 0 for median home value. These tracts were examined in the 2014-2018 ACS and 2010 Census and finding ‘0’ values there as well, so the median home value variable was removed in this study. Table 3 lists the remaining final variables used in the exploratory and generalized linear regression analysis.

Table 3 Final demographic variables used in regression models

| Final Variables |
| --- |
| 2018 Household Income $200,000 or greater (HINC200_CY_N) |
| 2018 Household Income less than $15,000 (HINC0_CY_N) |
| 2018 Median Household Income (MEDHINC_CY) |
| 2018 Owner-Occupied Housing Units (OWNER_CY_N) |
| 2018 Renter Occupied Housing Units (RENTER_CY_N) |
| 2018 Vacant Housing Units (VACANT_CY_N) |
| 2018 Male Population (MALES_CY_N) |
| 2018 Female Population (FEMALES_CY_N) |
| 2018 Unemployment Rate (UNEMPRT_CY) |
| 2018 Diversity Index (DIVINDX_CY) |

An exploratory regression analysis was repeated using the remaining variables. Summaries of the variable significance and multicollinearity were generated (Table 4). These summaries were used to understand what the variables were doing in the models. There were no violations for multicollinearity except for the male population variable at the census tract scale. The VIF value of 8.08 was slightly higher than the preferred standard of 7.5 in this instance. However, the male population variable was kept in the regression analysis model because the VIF violation was not present at the census block group scale. This is another indication that regression analysis is sensitive to changes in the scale used for analysis and suffers from the MAUP. Included in the summary of variable significance, is a significance rating on a scale of 0-100 for each candidate variable, and whether the linear relationship of the variable to property crime is primarily positive or negative. The higher the variable significance rating, the stronger the variable is a predictor of property crime. Variable significance percentages and order changed from census block group to census tract. The percentage value of whether a variable is negatively or positively related to property crime also changed from census block group to census tract.

Table 4 Census block groups & tracts exploratory regression summaries

Summary of Variable Significance

| Variable | % Significant | % Negative | % Positive |
|---|---|---|---|
| RENTER_CY_N | 100.00 | 0.00 | 100.00 |
| DIVINDX_CY | 80.08 | 3.78 | 96.22 |
| FEMALES_CY_N | 78.69 | 78.09 | 21.91 |
| MALES_CY_N | 72.71 | 12.55 | 87.45 |
| HINC0_CY_N | 65.54 | 0.00 | 100.00 |
| HINC200_CY_N | 60.36 | 0.80 | 99.20 |
| VACANT_CY_N | 53.98 | 0.00 | 100.00 |
| UNEMPRT_CY | 48.80 | 92.23 | 7.77 |
| MEDHINC_CY | 42.43 | 38.65 | 61.35 |
| OWNER_CY_N | 1.20 | 43.82 | 56.18 |

Summary of Multicollinearity

| Variable | VIF | Violations | Covariates |
|---|---|---|---|
| UNEMPRT_CY | 1.26 | 0 | -------- |
| DIVINDX_CY | 1.63 | 0 | -------- |
| MEDHINC_CY | 2.17 | 0 | -------- |
| OWNER_CY_N | 1.18 | 0 | -------- |
| RENTER_CY_N | 2.75 | 0 | -------- |
| VACANT_CY_N | 1.52 | 0 | -------- |
| HINC0_CY_N | 1.66 | 0 | -------- |
| HINC200_CY_N | 1.39 | 0 | -------- |
| MALES_CY_N | 6.37 | 0 | -------- |
| FEMALES_CY_N | 5.37 | 0 | -------- |

Summary of Variable Significance

| Variable | % Significant | % Negative | % Positive |
|---|---|---|---|
| RENTER_CY_N | 100.00 | 0.00 | 100.00 |
| FEMALES_CY_N | 84.86 | 92.03 | 7.97 |
| DIVINDX_CY | 74.50 | 7.17 | 92.83 |
| MEDHINC_CY | 70.72 | 14.14 | 85.86 |
| HINC0_CY_N | 66.73 | 18.13 | 81.87 |
| MALES_CY_N | 61.16 | 24.90 | 75.10 |
| VACANT_CY_N | 50.80 | 0.00 | 100.00 |
| OWNER_CY_N | 36.25 | 81.87 | 18.13 |
| UNEMPRT_CY | 25.10 | 74.50 | 25.50 |
| HINC200_CY_N | 17.93 | 40.64 | 59.36 |

Summary of Multicollinearity

| Variable | VIF | Violations | Covariates |
|---|---|---|---|
| UNEMPRT_CY | 1.78 | 0 | -------- |
| DIVINDX_CY | 1.87 | 0 | -------- |
| MEDHINC_CY | 4.84 | 0 | -------- |
| OWNER_CY_N | 2.12 | 0 | -------- |
| RENTER_CY_N | 4.42 | 0 | -------- |
| VACANT_CY_N | 2.38 | 0 | -------- |
| HINC0_CY_N | 5.90 | 0 | -------- |
| HINC200_CY_N | 2.36 | 0 | -------- |
| MALES_CY_N | 8.08 | 57 | -------- |
| FEMALES_CY_N | 7.15 | 0 | -------- |

Table 4 summarizes the differences in the exploratory regression results at each scale.

The renter-occupied housing units variable was the only variable where no changes occurred for the variable significance, positive, and negative percentages. The renter-occupied housing units variable was 100% significant, with a 100% positive relationship to property crime at the census block group and census tract scales. Two other variables, the vacant housing units and household income less than $15,000 variables retained the same rank in terms of significance. The household income ≥ $200,000 variable saw a large change in significance rank, significance percentage, and relationship percentage from the census block group to census tract. The rank decreased from 6 to 10, and the percent significance decreased from 60.36 to 17.93%. The largest change in significance was seen in the owner-occupied housing units variable. The variable at the census block group came in last at 1.20% significant, but it rose to 8$^{th}$ position, with a significance of 36.25%. The linear relationship of the variable changed from 43.82%

negative and 56.18% positive at the census block group to 81.87% negative and 18.13% positive at the census tract.

Table 5 Exploratory regression results at the census block group and census tract scales

| Variable | Rank | % Significant | % Negative | % Positive |
|---|---|---|---|---|
| Renter Occupied Housing Units (BG) | 1 | 100 | 0 | 100 |
| Renter Occupied Housing Units (Tract) | 1 | 100 | 0 | 100 |
| Diversity Index (BG) | 2 | 80.08 | 3.78 | 96.22 |
| Diversity Index (Tract) | 3 | 74.5 | 7.17 | 92.83 |
| Female Population (BG) | 3 | 78.69 | 78.09 | 21.91 |
| Female Population (Tract) | 2 | 84.86 | 92.03 | 7.97 |
| Male Population (BG) | 4 | 72.71 | 12.55 | 87.45 |
| Male Population (Tract) | 6 | 61.16 | 24.9 | 75.1 |
| Household Income < $15,000 (BG) | 5 | 65.54 | 0 | 100 |
| Household Income < $15,000 (Tract) | 5 | 66.73 | 18.13 | 81.87 |
| Household Income ≥ $200,000 (BG) | 6 | 60.36 | 0.8 | 99.2 |
| Household Income ≥ $200,000 (Tract) | 10 | 17.93 | 40.64 | 59.36 |
| Vacant Housing Units (BG) | 7 | 53.98 | 0 | 100 |
| Vacant Housing Units (Tract) | 7 | 50.8 | 0 | 100 |
| Unemployment Rate (BG) | 8 | 48.8 | 92.23 | 7.77 |
| Unemployment Rate (Tract) | 9 | 25.1 | 74.5 | 25.5 |
| Median Household Income (BG) | 9 | 42.43 | 38.65 | 61.35 |
| Median Household Income (Tract) | 4 | 70.72 | 14.14 | 85.86 |
| Owner Occupied Housing Units (BG) | 10 | 1.2 | 43.82 | 56.18 |
| Owner Occupied Housing Units (Tract) | 8 | 36.25 | 81.87 | 18.13 |

### *4.2.2. Descriptive Statistics of Final Variables*

Descriptive statistics show the differences between the variables across the two scales (Table 6). Descriptive statistics were run on the final variables. The variable format used for the descriptive statistics were counts, not the normalized data, except for the property crime, median household income, unemployment rate, and Diversity Index variables. The crime variable was a percent of all counts.

Table 6 Descriptive statistics of model variables

| Variable | Units | Mean | St. D | Min | Max |
|---|---|---|---|---|---|
| Crime Percent (BG) | ratio | 0.173611 | 0.338892 | 0.006825 | 4.886955 |
| Crime Percent (Tract) | ratio | 0.520833 | 0.767612 | 0.029008 | 5.895401 |
| Diversity Index (BG) | rank | 62.92083 | 14.42841 | 6.3 | 92.1 |
| Diversity Index (Tract) | rank | 64.37135 | 14.42762 | 20.5 | 92.2 |
| Median Household Income (BG) | ratio | 92732.46 | 36972.66 | 10714 | 200001 |
| Median Household Income (Tract) | ratio | 89305.35 | 34586.07 | 12734 | 198062 |
| Unemployment Rate (BG) | ratio | 3.602778 | 2.957789 | 0 | 20.2 |
| Unemployment Rate (Tract) | ratio | 3.76875 | 2.216847 | 0.2 | 15.9 |
| Owner-Occupied Housing Units (BG) | count | 40.76605 | 23.22956 | 0 | 92.37288 |
| Owner-Occupied Housing Units (Tract) | count | 37.67807 | 21.82971 | 0 | 85.26104 |
| Renter Occupied Housing Units (BG) | count | 52.28878 | 22.42109 | 19 | 4126 |
| Renter Occupied Housing Units (Tract) | count | 55.3487 | 20.92732 | 26 | 6538 |
| Vacant Housing Units (BG) | count | 6.94517 | 4.393559 | 0.522778 | 40.93366 |
| Vacant Housing Units (Tract) | count | 6.973231 | 3.958612 | 1.151493 | 37.37575 |
| Male Population (BG) | count | 50.38699 | 4.858465 | 36.66667 | 75.46012 |
| Male Population (Tract) | count | 50.62299 | 5.051038 | 39.52711 | 75.46012 |
| Female Population (BG) | count | 49.61301 | 4.858465 | 24.53988 | 63.33333 |
| Female Population (Tract) | count | 49.37701 | 5.051038 | 24.53988 | 60.47289 |
| Household Income < $15,000 (BG) | count | 9.370809 | 10.1766 | 0 | 70 |
| Household Income < $15,000 (Tract) | count | 10.39659 | 9.943143 | 1.117686 | 58.89952 |
| Household Income ≥ $200,000 (BG) | count | 0 | 12.35556 | 0 | 64.53901 |
| Household Income ≥ $200,000 (Tract) | count | 0 | 10.95981 | 0 | 49.69072 |

*4.2.3. Generalized Linear Regression Results*

Generalized linear regression (GLR) uses the same ordinary least square regression method as exploratory regression but differs in how the results are provided. The application generalized linear regression at the census block group and census tract scales resulted in differences in the coefficient linearity, probability, statistical significance, VIF values, AIC values, and adjusted $R^2$ values (Tables 7 and 8). Variable coefficient linearity changed for three of the 10 variables, and the coefficient strengths were altered for all 10 of the variables. The unemployment rate variable was negative at the census block group scale, and positive at the

census tract scale (Table 8). The owner-occupied housing units variable was positive at the census block group scale and negative at the census tract scale. The household income ≥ $200,000 variable was positive at the census block group scale and negative at the census tract scale. The changes in linearity can be caused by changes in the values of the aggregated data, as seen in the exploratory regression analysis results (Table 5). The VIF values increased as the aggregation unit increased from census block groups to census tracts (Table 8). An asterisk indicates the probability and robust probability (Robust_Pr) are statistically significant, but the level of significance also changed when the scale changed. The vacant housing units variable at the census tract scale had a statistically significant probability, but not at the census block group scale. The household income ≥ $200,000 variable had a statistically significant probability and robust probability at the census tract scale but not at the census block group scale.

Table 7 Generalized linear regression model diagnostics

| | | | |
|---|---|---|---|
| Number of Observations (BG): | 576 | Akaike's Information Criterion (AICc): | 1293.231623 |
| Number of Observations (Tract): | 192 | Akaike's Information Criterion (AICc): | 381.637326 |
| Multiple R-Squared (BG): | 0.485232 | Adjusted $R^2$: | 0.476121 |
| Multiple R-Squared (Tract): | 0.542166 | Adjusted $R^2$: | 0.516871 |
| Joint F-Statistic (BG): | 53.25818 | Prob(>F), (10,565) degrees of freedom: | 0.000000* |
| Joint F-Statistic (Tract): | 21.433965 | Prob(>F), (10,181) degrees of freedom: | 0.000000* |
| Joint Wald Statistic (BG): | 533.658364 | Prob(>chi-squared), (10) degrees of freedom: | 0.000000* |
| Joint Wald Statistic (Tract): | 248.781098 | Prob(>chi-squared), (10) degrees of freedom: | 0.000000* |
| Koenker (BP) Statistic (BG): | 41.024326 | Prob(>chi-squared), (10) degrees of freedom: | 0.000011* |
| Koenker (BP) Statistic (Tract): | 25.539279 | Prob(>chi-squared), (10) degrees of freedom: | 0.004412* |
| Jarque-Bera Statistic (BG): | 174.322857 | Prob(>chi-squared), (2) degrees of freedom: | 0.000000* |
| Jarque-Bera Statistic (Tract): | 25.21975 | Prob(>chi-squared), (2) degrees of freedom: | 0.000003* |

The regression models' $R^2$ values increased from 48% at the census block group scale to 54% at the census tract scale (Table 7). The models' adjusted $R^2$ values, a better indicator of the model performance than the multiple $R^2$ values, increased from 47% for the census block groups

to 51% for the census tracts. The differences between the models are due to the scale effect of the MAUP.

Table 8 Generalized linear regression results

| Variable | Coefficient | Probability | Robust_Pr | VIF |
|---|---|---|---|---|
| Intercept (BG) | -4.175816 | 0.000000* | 0.000118* | |
| Intercept (Tract) | -4.333853 | 0.000000* | 0.000039* | |
| Unemployment Rate (BG) | -0.000099 | 0.993267 | 0.992895 | 1.260365 |
| Unemployment Rate (Tract) | 0.020821 | 0.451028 | 0.462967 | 1.799194 |
| Diversity Index (BG) | 0.013802 | 0.000001* | 0.000026* | 1.640185 |
| Diversity Index (Tract) | 0.017221 | 0.000107* | 0.000262* | 1.879304 |
| Median Household Income (BG) | 0.000004 | 0.000740* | 0.002166* | 2.173575 |
| Median Household Income (Tract) | 0.000014 | 0.000005* | 0.000005* | 4.840523 |
| Owner-Occupied Housing Units (BG) | 0.054556 | 0.054649 | 0.303144 | 1.185458 |
| Owner-Occupied Housing Units (Tract) | -0.018501 | 0.732632 | 0.796711 | 2.119452 |
| Renter Occupied Housing Units (BG) | 0.716471 | 0.000000* | 0.000000* | 2.78653 |
| Renter Occupied Housing Units (Tract) | 0.636554 | 0.000001* | 0.000001* | 4.435324 |
| Vacant Housing Units (BG) | 0.088075 | 0.06783 | 0.166905 | 1.518525 |
| Vacant Housing Units (Tract) | 0.206148 | 0.036685* | 0.104551 | 2.384923 |
| Household Income < $15,000 (BG) | 0.081308 | 0.009277* | 0.034546* | 1.661251 |
| Household Income < $15,000 (Tract) | 0.342867 | 0.005780* | 0.002463* | 5.926499 |
| Household Income ≥ $200,000 (BG) | 0.012074 | 0.627874 | 0.699758 | 1.399957 |
| Household Income ≥ $200,000 (Tract) | -0.138725 | 0.011745* | 0.015300* | 2.377661 |
| Male Population (BG) | 1.105138 | 0.000000* | 0.000003* | 6.370187 |
| Male Population (Tract) | 0.931577 | 0.001237* | 0.004860* | 8.088657 |
| Female Population (BG) | -1.464225 | 0.000000* | 0.000000* | 5.372064 |
| Female Population (Tract) | -1.281492 | 0.000001* | 0.000068* | 7.210129 |

The generalized linear regression tool provided output feature classes of the standardized residuals for the model at the census block group and census tract scales. The standardized residuals show where the model is over- and under-predicting the dependent variable based on the observed values. Residuals compare the observed values to the predicted values on a linear prediction line (Figure 12). Observed values less than the predicted values result in over-prediction, observed values more than the prediction line result in under-prediction. The census

block group and census tract models show both types of predictions (Figures 15 and 16). The over- and under-predictions also suggest that key variables are missing in the study area that would help to explain property crime occurrence. Missing could be commercial areas, business centers, tourist spots, recreation or nightlife areas, police departments.

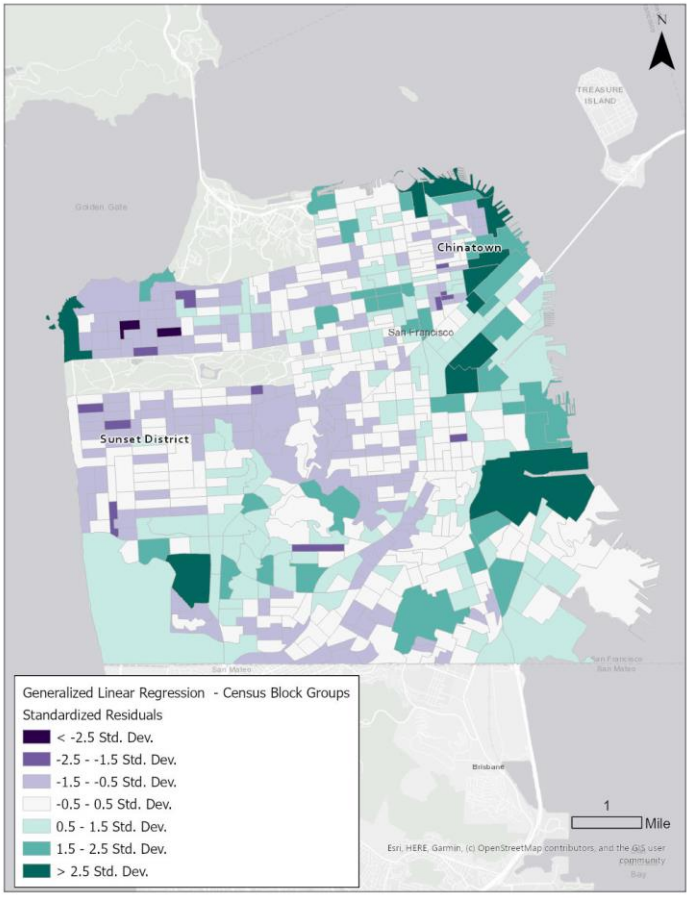


Figure 15 GLR census block group standardized residuals

When comparing the standardized residuals for the census block groups and census tracts, the results were different. The most significant difference was in the southwest. For the census block groups (Figure 15), the standardized residuals are below the mean, and for the census tracts, they are above the mean (Figure 16). Differences in standardized residuals for the census tracts and census block groups in the same locations in the study area suggest the regression analysis susceptibility to the MAUP scale effect.

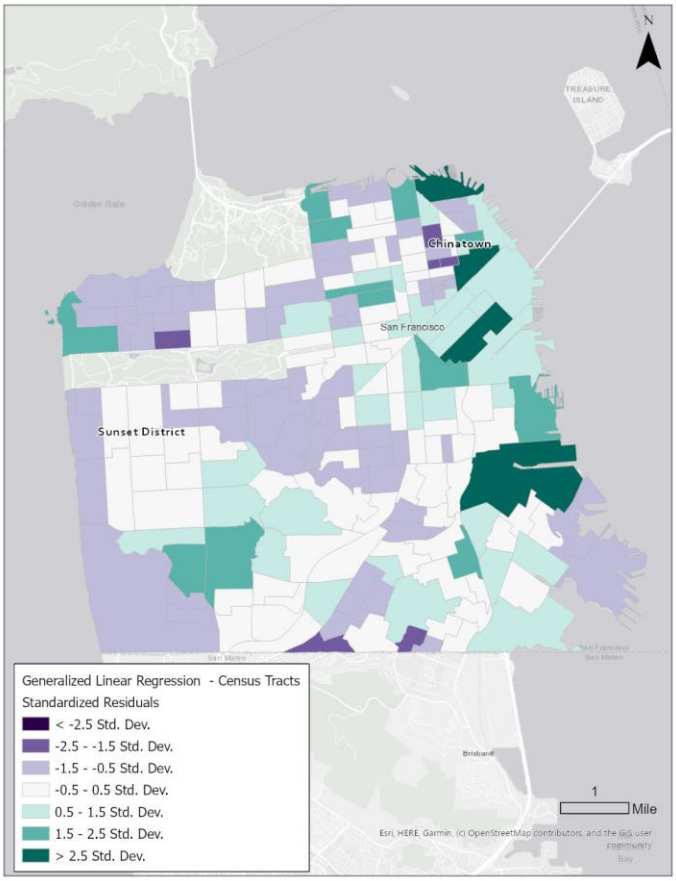


Figure 16 GLR census tract standardized residuals

The census block group GLR residuals present more of a heterogeneous pattern than the more homogenous census tracts. Although the regression models used the same datasets and variables, they produced different results due to the different scales of analysis. The MAUP causes the differences. One of the MAUP effects is that variables become more correlated as the aggregation level increases. When comparing the two GLR residual maps, the standardized residuals for the census tracts present a more homogenous pattern than the census block groups, and the increased correlation of the variables can explain this result. The patterns of the residuals also indicate that the variable relationships may not stationary throughout the study area.

# Chapter 5 Discussion and Conclusions

As introduced in Chapter 1, the MAUP will be a problem when using data that has been aggregated to aerial units or partitioned to zones. These are two ways the MAUP presents itself, as the MAUP scale effect and the MAUP zone effect. The zone effect emerges when data is processed on the same scale, but the delineation of the data changes. The scale effect, which this thesis focused on, occurs when data is aggregated to aerial units that have different scales. The selection level for aggregation impacts the visualization of the data. Data depicted using census tracts versus census block groups or counties versus states can present very different results when using the same underlying data. The goal of this study was not to show how the aggregation of data at different scales changes how information is visualized. The goal was to take the issue of the MAUP one step further and evaluate the scale effects of the MAUP on statistical analysis of property crime occurrence in San Francisco, California. The research question, which asked whether the optimized hot spot analysis and generalized linear regression results were different at the census block group and census tract scales due to the MAUP, was confirmed.

The findings of this study are similar to the findings of prior studies. For example, Fotheringham and Wong (1991) found that the aggregation of data causes a smoothing effect and results in a decrease in the variation evident in the data. The smoothing effect is more than likely what happened to the property crime as the scale of aggregation was increased from the census block group to the census tract. As the scale of aggregation increased, the local relationships and dynamics of the data were lost, leading to higher data heterogeneity.

The MAUP scale effects resulted in optimized hot spot analysis results that were different at the census block group and census tract scales. At the census block group scale, there were hot

spots and cold spots present that were not present at the census tract scale. The MAUP scale effect on optimized hot spot analysis is due to the change in the counts of the features and how distances are calculated for feature neighbors. The fixed distance band determines the optimal scale of analysis, and this study used. When the average nearest neighbors changed because of the change of scale, the distance to significant peaks in data clusters also changed. When data is aggregated to different boundaries, the weighted values also change because the value for each feature is compared proportionally to the sum of all features. If the value is more than the sum, the process at that location is not considered random, and that feature is given a statistically significant z-score. The effects of the weighted features and distance band thresholds for the Getis-Ord Gi* statistic were seen in the results of the optimized hot spot analysis.

The weight of each feature, based on the sum of the crime points, had the biggest effect on the MAUP. As aggregation increased from census block groups to census tracts, and the crime counts in local neighborhoods were combined, the variances decreased, causing low counts to increase or high counts to decrease. The changes resulted in the smoothing effect in the data due to aggregation described by Fotheringham and Wong (1991). As a result of the smoothing effect, the cold spot present at the census block group, disappeared at the census tract scale, and the hot spot areas decreased in terms of the geographic extent. For the census block groups, 17% of the features were considered statistically significant, and for the census tracts, 8.3% of the features were considered statistically significant. The statistically significant hot and cold spot areas decreased by 2.5 square miles for the census tracts. These findings are similar to those in McKay (2018). McKay conducted hot spot analysis of crime in Strathclyde at two different scales: output areas and data zones. The data zones are nested within the output areas.

The hot spot analysis had more areas that were colds spots at the data zone scale, and a large statistically significant hot spot appeared that was not present at the data zone scale.

The MAUP effects on hot spot analysis, depend on the data, scale, weight or count of each feature, and the number of neighbors each feature has, which is determined by the distance band threshold. The appearance or disappearance of statistically significant hot or cold spots can occur as the scale increases or decreases. The cluster analyses highlight anomalies in the data are present. If those anomalies change with scale, as demonstrated by the property crime in this study, efforts to addresses property crime, such as implementing safety measures, may not be effective. It is important to note that cluster analyses may also be affected by the study area shape and size. For example, in this study, the coastline and removed census tracts (Golden Gate Park and Presidio) were physical barriers and impacted how the fixed distance band and neighbors were calculated.

Statistical analysis sensitivity to the MAUP was also seen in the regression analysis results. The exploratory regression was used to determine the final selection of the variables for the GLR model. Once the final selection of variables was determined, exploratory regression was conducted again, to evaluate the significance of the variables as predictors of property crime occurrence at each scale. The results were not consistent at the two aggregation scales. The variables significance as predictors of property crime changed. Changes were seen in the significance percentage and rank between the variables at the census block group and census tract scales. The positive and negative linear relationships of some of the variables to property crime also changed with scale.

Scale changes were also seen in the model coefficients, which indicate the strength and relationship of each variable to property crime, and the variable coefficient probabilities, which

indicate whether a coefficient is statistically significant in the final GLR models. The three variables that had coefficients change from positive at one scale to negative at another scale were the unemployment rate, the number of owner-occupied housing units, and household income ≥ $200,000. Statistical significance changes in the coefficient probabilities were seen for the household income ≥ $200,000 and the number of vacant housing units. The GLR models adjusted $R^2$ values also changed. These values are an indicator of the model performance. The adjusted $R^2$ value was higher for the census tracts (51%) than census block groups (47%). The increased adjusted $R^2$ squared values confirmed earlier studies which found that correlations among variables strengthen as data is aggregated due to the MAUP (Gehlke and Biel, 1934; Openshaw and Taylor, 1979; Wong, 2009; Fotheringham and Wong, 1991; Cheng and Fotheringham, 2013; Saib et al. 2014).

The MAUP will always be a factor when data is aggregated to different scales or zones and should always be considered in spatial analysis. Therefore, a spatial analysis should be conducted at multiple scales if the data is available to do so. Decisions or assumptions made based on the analytical results using aggregated data could be wrong. This is the risk of not analyzing data at multiple scales to determine the presence and severity of the MAUP.

The concepts and methods used in this study provide a framework for evaluating the MAUP effects, even if researchers use different data and tools. Although in many GIS core curricula, the MAUP is regarded as a core concept, many professionals and researchers often fail to consider it in their analysis. One of the aims of this study is to raise awareness of the MAUP among people who conduct spatial analysis.

## 5.1. Limitations and Recommendations

### *5.1.1. Assessing Model Significance to Property Crime in San Francisco*

If evaluating the model significance to property crime in San Francisco, the variables used for the regression models in this study explain approximately 50 % of property crime occurrence. Due to this, a further avenue of research could be identifying and using variables that are specific to just the routine activity theory or the social disorganization theory. In addition, crime theory models often combine two or more variables to create a new variable. For example, combining two variables to create a variable measuring disadvantage or poverty in an area. This study did not generate combined variables, which is another possible reason why the model only explains about half of the property crime occurrence in San Francisco.

### *5.1.2. Future Work and Avenues for Research*

One avenue for future research would be to evaluate the same outcome using GWR. A local form of regression, such as GWR, may lessen the effect of the MAUP. Global regression models assume that the relationships are the same over the whole study area, where, GWR takes into consideration that the relationships may vary over space and calculates regression models for each feature. OLS is a global form of regression and assumes the relationships are static and consistent over space and uses a single equation for the study area. GWR would result in a better model because it reflects Tobler's First Law of Geography. However, Cheng and Fotheringham (2013) confirmed that the MAUP is still present in the results of GWR, but less so compared to OLS.

The second avenue for future research would be to conduct bivariate correlation analysis to test the model variables at each scale. Bivariate correlation analysis is used to model the relationships between the dependent and independent variables, and well as the relationships of

independent variables to each other. Bivariate correlation analysis models relationships that are nonlinear, whereas the GLR and exploratory regression analysis use ordinary least squares, which assumes that the relationships are linear. The regression residuals patterns from the GLR in this study indicate that variable relationships to property crime may not be linear or standard over the study area. Bivariate correlation analysis can assist in exploring spatial autocorrelation and non-stationarity between the variables.

The third avenue for research is adding variables that may contribute to crime besides sociodemographic factors. A possible reason the model explains about 50% of the variability of property crime, could be that the model did not consider some key variables. Examples of key variables could be missing spatial features like stores, police stations, tourism and nightlife attractions and business districts. In addition to the missing spatial features, another analysis could be done to explore the property crime analysis not in just space, but in time. There are certain times of day where areas or people within those areas may make opportune targets.

The fourth avenue for future research is to conduct regression and cluster analysis by using normalized and non-normalized data. Pietrzak (2014) and Louvet et al. (2015) explored the impact of the MAUP in such a manner. They suggest that normalized data is less susceptible to the scale effects of the MAUP. The cluster analysis performed in this study used only property crime counts. Therefore, evaluating the MAUP scale effects on the results of the cluster analysis by comparing cases of normalized (i.e., counts per unit acre or 1,000 residents) and non-normalized data (counts) might yield better results.

## 5.2. Final Thoughts

This study is relevant to assessing property crime and the MAUP scale effects. The MAUP is a well-known issue in the spatial sciences; however, non-GIS professionals, crime

analysts, and government officials may not be familiar with the matter. This study highlights how the MAUP applies to multiple disciplines, increases MAUP awareness, and provides a framework that can be replicated in other studies.

Crime analysts and public stakeholders in San Francisco, and other cities, can use the methods in this study when addressing and implementing safety measures to mitigate crime. As seen with the analyses in this study, identifying areas with significant property crime occurrence and what factors explain the phenomenon are affected by the MAUP scale effects. Therefore, analyzing crime at multiple scales should also be considered because areas of interest are subject to change as the scale changes.

Stakeholders must understand assessments of causative crime attractors fluctuate with scale due to the MAUP. For example, the unemployment rate variable was negatively related to property crime at the census block group scale and positively related at the census tract scale. Why was the unemployment rate contributary to property crime at the census tract scale but not at the census block group scale? The household income ≥ $200,000 variable was positively related to crime at the census block group scale and negatively related at the census tract scale. Are households with higher incomes only targets for property crime in areas that are not surrounded by areas with higher household incomes? The owner-occupied housing units variable was negative at the census tract scale, and positive at the census block group scale. These results lead to different assumptions and subsequent questions at each scale. Analysts and researchers conducting spatial analyses examining socio-demographic phenomena such as crime must closely scrutinize resultant data in terms of the MAUP.

# References


Andresen, Martin A. 2006. "A Spatial Analysis of Crime in Vancouver, British Columbia: a Synthesis of Social Disorganization and Routine Activity Theory." *Canadian Geographer / Le Géographe canadien* 50, no. 4 (December): 487–502.

Arbia, Giuseppe, and Petrarca, Francesca. 2011. "Effects of MAUP on Spatial Econometric Models." *Letters in Spatial and Resource Sciences* 4, no. 3 (October): 173–185.

Bolstad, Paul. 2016. *GIS Fundamentals: A First Text on Geographic Information Systems* (5th Edition). Ann Arbor, MI: XanEdu Inc.

Cahill, Meagan, and Gordon Mulligan. 2007. "Using Geographically Weighted Regression to Explore Local Crime Patterns." *Social Science Computer Review* 25, no. 2 (May): 174–193.

Cassidy, Meghan and Ravani, Sarah. 2018. "The Scanner: San Francisco ranks No. 1 in US in property crime". October 1. https://www.sfchronicle.com/crime/article/The-Scanner-San-Francisco-ranks-No-1-in-13267113.php. (accessed October 1, 2019).

Cheng, Jianquan, and Fotheringham, A. Stewart. 2013. "Multi-Scale Issues in Cross-Border Comparative Analysis." *Geoforum* 46, no. C: 138–148.

Dailey, George. 2006. "Normalizing Census Data Using ArcMap". *Arc User*. (January-March): 52-53

Duque, Juan. C., Laniado, Henry, and Polo, Adriano. 2018. "S-Maup: Statistical Test to Measure the Sensitivity to the Modifiable Areal Unit Problem." *PLoS ONE* 13, no. 11: e0207377.

Esri. 2012. "Vendor Accuracy Study: 2010 Estimates versus Census 2010". Esri Demographics. http://www.esri.com/~/media/Files/Pdfs/library/brochures/pdfs/vendor-accuracy-study.pdf

Esri. 2018. "Methodology Statement: 2018/2023 Esri US Updated Demographics". An Esri White Paper. https://doc.arcgis.com/en/esri-demographics/data/updated-demographics.htm#GUID-258B26CB-A833-4B85-97F5-2DD664FAD1D0

Esri. 2019a. "How Hot Spot Analysis (Getis-Ord Gi*) works". ArcGIS Pro help. https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/h-how-hot-spot-analysis-getis-ord-gi-spatial-stati.htm. (accessed November 1, 2019).

Esri. 2019b. "Ordinary Least Squares (OLS) (Spatial Statistics). ArcGIS Pro help. https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/ordinary-least-squares.htm (accessed March, 9, 2020).

Esri. 2019c. "Regression analysis basics". ArcGIS Pro help. https://pro.arcgis.com/en/pro-app/tool-reference/spatial-statistics/regression-analysis-basics.htm. (accessed November 1, 2019).

Flowerdew, Robin. 2011. "How Serious Is the Modifiable Areal Unit Problem for Analysis of English Census Data?" *Population Trends*, no. 145 (October): 106–106. http://search.proquest.com/docview/1093463886/.

FBI Uniform Crime Reporting Program. 2017. Crime in the United States. https://ucr.fbi.gov/crime-in-the-u.s/2017/crime-in-the-u.s.-2017/topic-pages/offense-definitions (accessed December 1, 2019).

Fotheringham, A.S, and Wong, D W S. 1991. "The Modifiable Areal Unit Problem in Multivariate Statistical Analysis." *Environment and Planning* A 23, no. 7 (July): 1025–1044.

Fotheringham, A. Stewart, Yang, Wenbai, and Kang, Wei. 2017. "Multiscale Geographically Weighted Regression (MGWR)." *Annals of the American Association of Geographers* 107, no. 6 (November 2): 1247–1265. http://www.tandfonline.com/doi/abs/10.1080/24694452.2017.1352480.

Gehlke, C. E., and Katherine Biehl. 1934. "Certain Effects of Grouping Upon the Size of the Correlation Coefficient in Census Tract Material." *Journal of the American Statistical Association* 29, no. 185 (March 1): 169–170.

Hart, Timothy, and Waller, Jeremy. 2013. "Neighborhood Boundaries and Structural Determinants of Social Disorganization: Examining the Validity of Commonly Used Measures." *Western Criminology Review* 14, no. 3 (November 1): 16–33. http://search.proquest.com/docview/1503118201/.

Lee, Matthew, and Alshalan, Abdullah. 2005. "Geographic Variation in Property Crime Rates: A Test of Opportunity Theory." *Journal of Crime & Justice* (July 1): 101–127. http://search.proquest.com/docview/223880602/.

Lee, Youngmin, Kwon, Pil, Yu, Kiyun, and Park, Woojin. 2016. "Method for Determining Appropriate Clustering Criteria of Location-Sensing Data." *ISPRS International Journal of Geo-Information* 5, no. 9 (January 1): 151. http://search.proquest.com/docview/1819311414/.

Lloyd, Christopher D. 2014. Exploring Spatial Scale in Geography Chichester, England: Wiley-Blackwell.

Louvet, Romain, Jagannath Aryal, Didier Josselin, and Cyrille Genre-Grandpierre. 2015. "R as a GIS: Illustrating Scale and Aggregation Problems with Forest Fire Data." *Procedia Environmental Sciences* 27, no. C: 66–69.

Mackaness, William A. 2007. “Chapter 1 - Understanding Geographic Space.” *In Generalisation of Geographic Information*, 1–10. Elsevier Ltd

McKay, Rebecca M. 2018. Comparing Crime Hot spots at Different Areal Resolutions in Strathclyde. MSc Statistics Dissertation, University of Glasgow

Mennis, J. 2019. Problems of Scale and Zoning. The Geographic Information Science & Technology Body of Knowledge (1st Quarter 2019 Edition), John P. Wilson (Ed.). DOI: 10.22224/gistbok/2019.1.2

Mitchell, Andy. 2009. The Esri Guide to GIS Analysis, Vol. 2: Spatial Measurements and Statistics, Esri Press, Redlands, California.

Montello, Daniel. R., and Paul C. Sutton. 2013. *An introduction to scientific research methods in geography and environmental studies*. 2nd ed. Los Angeles, CA: Sage.

Moriarty, Laura, and Williams, James. 1996. “Examining the Relationship Between Routine Activities Theory and Social Disorganization: An Analysis of Property Crime Victimization.” *American Journal of Criminal Justice* 21, no. 1 (September): 43–59.

Nielsen, Michael M., and Hennerdal, Pontus. 2014. “MAUPing Workplace Clusters.” *Growth and Change* 45, no. 2 (June): 211–221.

Openshaw, S. 1977. “A geographical solution to scale and aggregation problems in region-building, partitioning and spatial modelling.” *Trans. Inst. Br. Geogr*. 2 (4):459-472.

Openshaw, S. and Taylor, P. J. (1979). A million or so correlation coefficients: Three experiments on the modifiable areal unit problem. In Wrigley, N. (ed.) Statistical Applications in the Spatial Sciences, pp 127--144. Pion: London.

Parenteau, Marie-Pierre, and Michael C. Sawada. 2011. “The Modifiable Areal Unit Problem (MAUP) in the Relationship Between Exposure to $NO_2$ and Respiratory health.” *International Journal of Health Geographics* 10, no.1 (October 31): 58.

Porter, Jeremy. 2011. “Identifying Spatio-Temporal Patterns of Articulated Criminal Offending: An Application Using Phenomenologically Meaningful Police Jurisdictional Geographies.” *Systems Research and Behavioral Science* 28, no.3 (May 1): 197-211. http://search.proquest.com/docview/872444854/.

Pietrzak, Michal Bernard. 2014. "The Modifiable Areal Unit Problem - Analysis of Correlation and Regression." *Equilibrium* 9 (4) (12): 113-131.

Saib, Mahdi-Salim, Caudeville, Julien, Carre, Florence, Ganry, Olivier, Trugeon, Alain, and Cicolella, Andre. 2014. “Spatial Relationship Quantification Between Environmental, Socioeconomic and Health Data at Different Geographic Levels.” *International journal of environmental research and public health* 11, no. 4 (April 3): 3765–3786.

San Francisco Government. 2018. *"Poverty in San Francisco."* City Performance Score Cards. https://sfgov.org/scorecards/safety-net/poverty-san-francisco (accessed 12/1/2019)

Swift, Andrew, Liu, Lin, and Uber, James. 2008. "Reducing MAUP Bias of Correlation Statistics Between Water Quality and GI Illness*." Computers, Environment and Urban Systems* 32, no2: 134–148.

Swift, Andrew, Liu, Lin, and Uber, James. 2014. "MAUP Sensitivity Analysis of Ecological Bias in Health Studies." *GeoJournal* 79, no. 2 (January 1): 137–153.

Tobler, W. 1970. "A Computer Movie Simulating Urban Growth in the Detroit Region." *Economic Geography* 46 (January 1). http://search.proquest.com/docview/1290146238/?pq-origsite=primo.

Wickes, Rebecca, Zahnow, Renee, Schaefer, Lacey, and Sparkes-Carroll, Michelle. 2017. "Neighborhood Guardianship and Property Crime Victimization." *Crime & Delinquency* 63, no. 5 (May): 519–544.

Wong, D.W. 2009. "Modifiable Areal Unit Problem." In International Encyclopedia of Human Geography, 169–174. Elsevier Ltd.

# Appendix A Optimized Hot Spot Analysis Results Windows

```
Running script OptimizedHotSpotAnalysis...
************************ Initial Data Assessment ************************
Making sure there are enough incidents for analysis....
- There are 58605 valid input features.
Evaluating the aggregation polygons....
- There are 576 valid input aggregation polygons.
Looking for locational outliers....
- There were 8 outlier locations; these will not be used to compute the optimal fixed
distance band.
************************** Incident Aggregation **************************
Counting the number of incidents in each polygon....
- Analysis is performed on all aggregation polygons.
Evaluating incident counts and number of polygons....
- The aggregation process resulted in 576 weighted polygons.
- Incident Count Properties:
        Min:          4.0000
        Max:       2864.0000
        Mean:       101.7448
        Std. Dev.:  198.4353
***************************** Scale of Analysis *****************************
Looking for an optimal scale of analysis by assessing the intensity of clustering at
increasing distances....
- No optimal distance was found using this method.
Determining an optimal distance using the spatial distribution of features....
- The optimal fixed distance band is based on the average distance to 28 nearest
neighbors: 4283.0000 US_Feet
***************************** Hot Spot Analysis *****************************
Finding statistically significant clusters of high and low incident counts....
- There are 99 output features statistically significant based on an FDR correction for
multiple testing and spatial dependence.
- 1% of features had less than 8 neighbors based on the distance band of 4283.0000
US_Feet
```

```
Running script OptimizedHotSpotAnalysis...
************************* Initial Data Assessment *************************
Making sure there are enough incidents for analysis....
- There are 58605 valid input features.
Evaluating the aggregation polygons....
- There are 192 valid input aggregation polygons.
Looking for locational outliers....
- There were 2 outlier locations; these will not be used to compute the optimal fixed
distance band.
************************** Incident Aggregation **************************
Counting the number of incidents in each polygon....
- Analysis is performed on all aggregation polygons.
Evaluating incident counts and number of polygons....
- The aggregation process resulted in 192 weighted polygons.
- Incident Count Properties:
        Min:         17.0000
        Max:       3455.0000
        Mean:       305.2344
        Std. Dev.:  448.6862
***************************** Scale of Analysis *****************************
Looking for an optimal scale of analysis by assessing the intensity of clustering at
increasing distances....
- No optimal distance was found using this method.
Determining an optimal distance using the spatial distribution of features....
- The optimal fixed distance band is based on the average distance to 9 nearest
neighbors: 4297.0000 US_Feet
***************************** Hot Spot Analysis *****************************
Finding statistically significant clusters of high and low incident counts....
- There are 16 output features statistically significant based on an FDR correction for
multiple testing and spatial dependence.
- 29.2% of features had less than 8 neighbors based on the distance band of 4297.0000
US_Feet
```