Residential Housing Code Violation Prediction:
A Study in Victorville, CA Using Geographically Weighted Logistic Regression

by

Matthew Dean Pugh

A Thesis Presented to the
Faculty of the USC Graduate School
University of Southern California
In Partial Fulfillment of the
Requirements for the Degree
Master of Science
(Geographic Information Science and Technology)

May 2016

I dedicate this paper to my loving parents, my brother, and to all of my family and friends who supported me throughout this entire process. Without you, none of this would have been possible.

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank my advisor, Dr. Karen Kemp, for her guidance and encouragement in completing this project. I would also like to thank my committee members, Dr. Robert Vos and Dr. Su Jin Lee, for their help in getting me to this point. I thank my parents, Allan and Donna, and my brother Steven for their constant support throughout my entire student career. Also, my friends Randy, Audrey, Bobby, Monique, Thomas and his wife Pilar, and Robbi for their support and understanding, and for reminding me to have a little fun from time to time. I also thank Robert Chang, PsyD for his sound advice and encouragement. I would also like to acknowledge my employers, the City of Victorville and Mimi Song Company for allowing me to use their resources to complete this project.

List of Abbreviations

ABC	Alcoholic Beverage Control
AIC	Akaike information criterion
BLR	Binary logistic regression
CD	Compact disk
GIS	Geographic information systems
GWR	Geographically weighted regression
GWLR	Geographically weighted logistic regression
IDRE	Institute for Digital Research and Education
NAD	North American Datum
OLS	Ordinary least squares
SFR	Single-family residence
SPSS	Statistical Package for Social Sciences
UCLA	University of California Los Angeles
USC	University of Southern California
VIF	Variance inflation factor

Abstract

Cities throughout the country are constantly striving to improve their perceived image. Whether it is requiring lush landscaping in commercial developments, or simply making sure that the trim on a house is properly painted, cities are constantly struggling to get citizens to comply with municipal codes. Such is the case in the City of Victorville, CA, where economic recovery has been slow following the 2008 housing market crash, leaving poorly maintained properties in its wake. Presently, Victorville's code enforcement staff is doing a proactive enforcement survey of all single-family homes in the city in an effort to "clean up" these properties. However, the survey is inefficient and is taking up a good amount of officer time, leaving commercial and industrial areas of the city neglected. This project was able to predict which houses in Victorville are likely to have a code enforcement violation that requires action from staff in order to better allocate resources to areas that require more attention and pull resources from areas that do not require attention. The primary question here is what property attributes can be used to predict the occurrence of a code enforcement violation? Several have been selected, including property value, length of ownership, and presence of a previous violation. A binary logistic regression analysis was run on three areas of the city containing approximately 2,200 homes that have already been surveyed in order to train a model for predicting the remaining 29,000 homes. Geographically weighted logistic regression was then employed to factor in spatial variation in the relationships between the response variable and the explanatory variables. The success of this model will make Victorville's code enforcement more efficient, and it is a model that any city can employ to make its own code enforcement departments more effective.

Chapter 1 Introduction

In the City of Victorville, which is situated in the Mojave Desert region of Southern California (Figure 1), there has been a complete shift in the quality of residential housing due to the construction boom of 2004 to 2008. Developers from all over came to the City and built low quality homes at a high volume in order to make the largest profit possible. Compounding this issue was a lack of planning and urban design that is essential in creating sustainable neighborhoods. Once the market crashed at the end of 2008, developers left town to seek their riches elsewhere. Left in their wake were architecturally basic, cheaply constructed single-family homes that are already showing signs of decay and blight. In the years that followed, property values remained low and a new wave of investors flooded the City to take advantage of homes that had been lost to bank foreclosures, short sales, and utter abandonment. The result was a high number of renter occupied homes that went unmaintained because occupants had no sense of place and owners of rental properties went unaccountable because they did not live in the area. This left front yards devoid of landscaping with weeds, deteriorating fences, inoperable vehicles, trash and debris, and an overall look of decay.

For these reasons, the City has adopted a new code enforcement policy of proactive enforcement of municipal laws that govern the aforementioned property maintenance issues in an effort to clean up the City, add some stability to residential districts, and allow neighborhoods to both retain and increase their property values. This has resulted in the hiring of three new code enforcement officers and an influx of new code enforcement action cases that take hours of manpower to rectify. In addition, as the seven total officers concentrate on residential properties, commercial and industrial properties get neglected. This creates inefficiency that the code enforcement manager and City Council need to address. This research project attempted to find

a better way of allocating those code enforcement resources by using statistical analysis techniques and a set of explanatory variables to predict which areas in the City are likely to have the greatest need for enforcement action. By doing this, City officials should be able to better manage their personnel, budget, and infrastructure in addition to providing a better level of service to the City's residents and businesses.

Figure 1: The City of Victorville

1.1 Motivation

The City's code enforcement division is currently conducting a citywide proactive investigation on single-family residential neighborhoods, logging which properties have housing code violations and which properties do not. Full neighborhood investigations are conducted and formal written notices of violation are being given to property owners or the occupant of the residence. The result is a set of notices with a property address that can be mapped easily at the parcel level; properties that do not have a violation are not given a notification. The code enforcement division has investigated roughly 2,300 single-family properties out of approximately 29,000 total single-family residences as of July 11, 2015, which is the date that the data used in this project was obtained from the City. The senior code enforcement officer estimates that this project will take another 18 to 24 months to complete, if not longer. This is putting a strain on officer time as they must also continue to address the other properties in the City where code enforcement action is necessary, such as in commercial or industrial areas. Making this process more efficient would greatly benefit the City as a whole and free up officer resources to fulfill all responsibilities for city code enforcement.

To do so, this project made predictions on which single-family residences are likely to have housing code violations and which are not likely to have housing code violations using the binary logistic regression technique. This technique is able to model binary outcomes such as yes/no, pass/fail, dead/alive, or 0/1. The model used various measures that either pertain to the property itself or to the neighborhood it is situated in and gave each property a value between 0 and 1 for where it fell on the logistic regression curve, or line of best fit. The result was a yes or no value depicting whether a code violation was likely or not. The resulting model can be used by the City's code enforcement staff to identify neighborhoods that should be given extra

attention or neighborhoods that will require very little attention. This should give staff a clear picture of the current state of the City and how they should proceed with their proactive inspection program to most efficiently manage the situation.

The hope is that this research project will extend beyond the scope of this analysis and be a great benefit to the City of Victorville. As mentioned before, the City is having great difficulty in keeping its residential housing stock in proper repair because of various economic and social issues. A successful regression model should be able to reliably predict areas where property maintenance is likely to be an issue, and the local government will be able to efficiently allocate its code enforcement resources to areas that will require the most attention. This will also give city officials a glimpse into the main contributing factors that cause residential properties to fall into decay, thus allowing officials to act accordingly. In addition, this project could help to alleviate the strain on property values from poor maintenance and hopefully raise residents' pride in their neighborhood and community. Furthermore, it could be used by other cities that are also struggling with their residential zones because the variables used in the analysis are not specific to Victorville. Instead, they are variables that are related to the characteristics of the homes themselves and could be obtained in any county or city that maintains property characteristic data and assessed value data.

To the best of my knowledge, binary logistic regression modeling has not been used to predict the occurrence of residential properties that require code enforcement action. Therefore, this project adds to previously conducted studies for prediction. Spatial predictive analysis has been conducted on crime (Antolos et al. 2013; Liu & Brown 2003), property abandonment (Morckel 2014), fire hazard potential (Rodrigues, de la Riva, & Fotheringham 2014; Martinez-Fernandez, Chuvieco, & Koutsias 2013) groundwater spring potential (Ozdemir 2011), riverbank

erosion potential (Atkinson et al. 2003), and landslide hazard potential (Kundu et al. 2013). This project gives another aspect to what the binary logistic regression technique can be used for in spatial analysis and prediction. Furthermore, this project will give researchers another look into how spatial regression techniques can be implemented in manners that effect public policy and decision-making for housing regulation.

In addition, there has been little research into how a binary logistic regression model can be applied to geographic data, and more specifically, how the occurrence of a dependent variable is affected by the location of another dependent variable in the dataset; in this case, a housing code violation. Therefore, this project adds to work done using geographically weighted regression with a binary outcome. Research conducted previously includes that of Atkinson et al. (2003) where they employed a geographically weighted logistic regression (GWLR) model for erosion susceptibility in order to study the effects of local phenomena that varied with location, such as distance upstream. In addition, Luo and Kanala (2008) employed GWLR to model urban growth and land use change over time. Also, the work of Martinez-Fernandez, Chuvievo and Koutsias (2013) accounted for local variations in wildfire occurrence variables using GWLR. This project will add to this previous research and will give an additional field of study for the GWLR technique to be employed beyond physical geography or sociology.

Moreover, the ability to predict other municipality related phenomena such as which properties are likely to become rental homes or which properties are likely to have crime could be profoundly useful to a local government. The techniques employed here could ultimately lead to other models that may be used by governing authorities to make their internal processes more efficient. This could lead to more man hours available for other projects and more budgetary funds to apply to other programs or departments. Even though this project is covering a very

small aspect of what local governments do, it potentially has the ability to provide lawmakers and officials with the ability to simplify their processes and provide a more effective form of government.

1.2 Research Questions

The goal of this project is to predict whether or not a single-family home in the City of Victorville is likely to have a code enforcement violation. This project will answer the following questions:

1. Can certain property attributes predict the occurrence of code enforcement violations?
2. Are there relationships between a home with a code enforcement violation and its neighboring properties?
3. Which single-family residential properties in Victorville are likely to have a violation?

Questions 1 and 2 will be answered as a result of the logistic regression and geographically weighted regression techniques. These questions are necessary because they essentially determine if this project has any validity in the field of geography. Question 1 will determine if the subsequent model is of any value scientifically and Question 2 will determine if spatial location does in fact play a role in how the physical characteristics of a neighborhood can be affected if there are a few properties that are substandard or run down. Question 3 is of course the goal of this research project.

1.3 Overview of Research Design

Since this model was predicting a binary outcome and some of the explanatory variables are non-continuous, techniques such as ordinary least squares and traditional geographically

weighted regression cannot be used. Following the examples of Lee and Sambath (2005), Wu and Zhang (2013), Martinez-Fernandez, Chuvieco, and Koutsias (2013), and Rodrigues, de la Riva, and Fotheringham (2014), a binary logistic and a geographically weighted logistic regression were created using various explanatory variables related to individual properties. Variables used were lot size, assessed land value, ownership type, length of ownership, year of construction, or whether or not a property is tax defaulted. The study focused on three specific areas within the City that have already been fully surveyed by code enforcement staff (see Figure 2). The binary logistic regression analysis was conducted on each of these areas independently. They were then combined into a single dataset where the analysis was conducted on a set of random samples from the dataset as well as on the data as a whole. The model was then analyzed for whether or not it violated any of the key assumptions of logistic regression, which are outlined later in this document.

The process was then repeated using only geographically weighted logistic regression within the GWR4 software developed by Tomoki Nakaya and distributed by Arizona State University (Nakaya 2009). The intent was to improve upon the logistic regression model by introducing spatial variation into the equation. This sought to illustrate that single-family residences can be affected by neighboring residences in regards to the need for code enforcement action. The GWR4 software produced variable coefficient values and a constant value that were then used in the geographically weighted logistic equation to make predictions for the remaining single-family homes in the City.

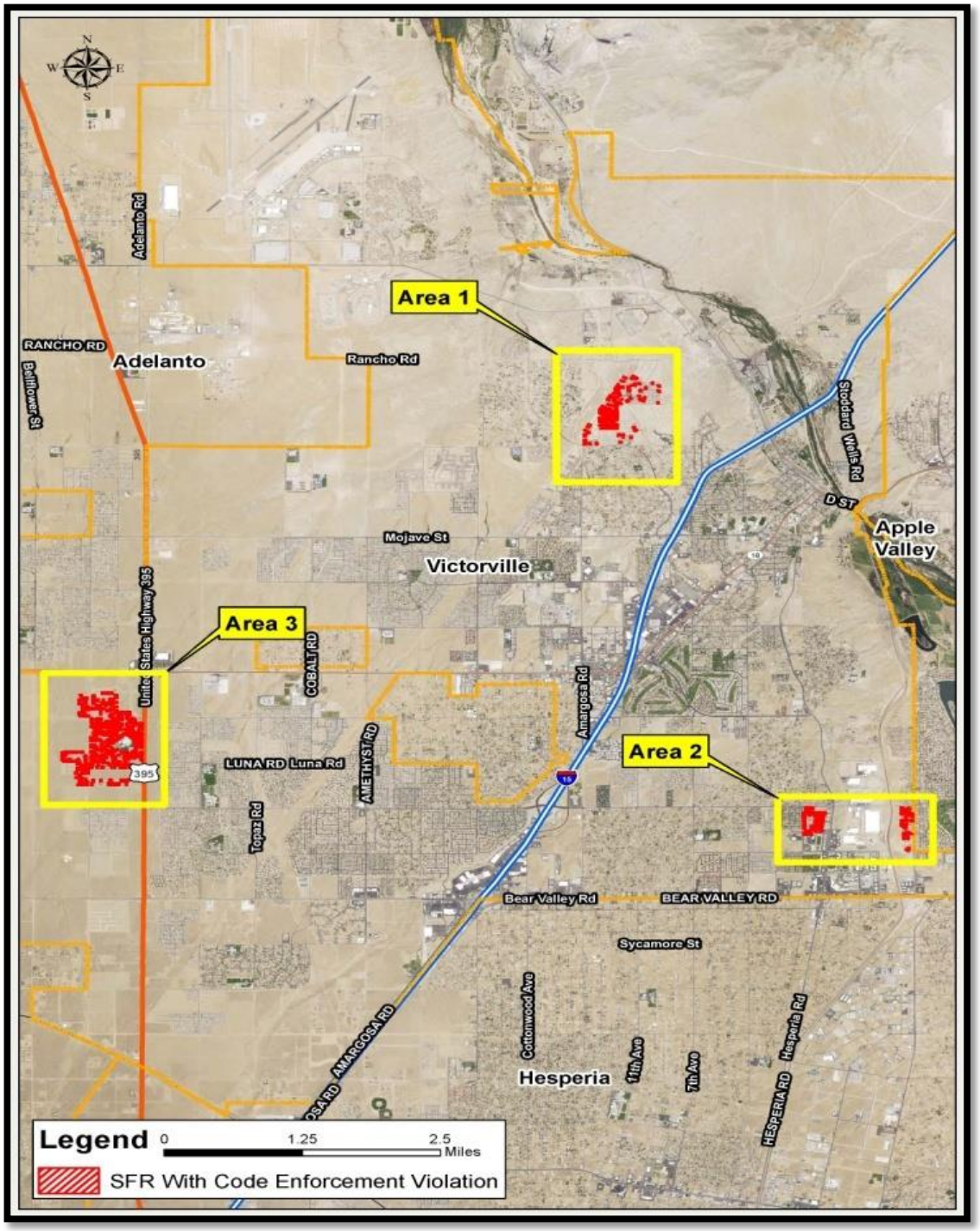


Figure 2: Study Areas 1, 2, and 3

The model was validated by evaluating the important assumptions of binary logistic regression and whether or not the model violated any of them. Next, the model was tested on an

area of the City that was proactively surveyed after the date of data collection of this project to determine how well the model performed as a predictive tool. The model performed well in terms of identifying neighborhoods that are likely to have violations, but it was not strong enough to confidently identify individual properties that are likely to have a violation.

The following chapters explain the details of how these predictions were made and the steps that were necessary to arrive here. Chapter 2 discusses the role of code enforcement and how property owners and neighborhoods are affected by code enforcement. It also discusses how crime incidents are predicted using geographic statistics because crime prediction has many similarities to code enforcement prediction, and it describes the applications of how logistic regression and geographically weighted logistic regression can be used as a prediction tool in the realm of geography. Chapter 3 outlines the methodology employed in creating this model, what data was needed, and how the predictions were made. Chapter 4 discusses the results of the analysis and how well the model performed as a predictive tool. Finally, Chapter 5 illustrates the limitations of this model, how it tied in to the related work that precedes it, and what can be done in the future to make this a stronger predictive tool.

Chapter 2 Background and Related Work

Cities in the United States have been dealing with the phenomena of residential housing decline, crime, and urban blight for decades. Generally, housing that was once new and attractive to potential residents slowly falls into a state of decay as homes or apartment buildings fall behind on general maintenance such as painting, landscape pruning and trimming, window replacement, roof replacement, or fence replacement. This, in conjunction with higher crime rates, gives the appearance that residential areas are dark, dangerous and no longer desirable places to live. As a result, cities must use their code enforcement resources to force residents to perform these essential property maintenance tasks to improve not only the neighborhood, but also the overall appearance of the city so that new residents will continue to move there, thus increasing the city's tax base and revenue (Anderson and Cordell 1988).

The City of Victorville is currently in this process of neighborhood revitalization through proactive code enforcement. This research project attempted to predict whether or not a property within the City would have some kind of code enforcement violation. In addition, an understanding of how the occurrence of an incident can be predicted through spatial analysis was essential. Crime prediction using spatial analysis was a good comparison because crime incidents and code enforcement incidents share similar attributes, such as they must be reported, they are point based, and they involve violation of a law. Next, various regression modeling techniques were analyzed to determine how prediction of incidents is possible through both traditional statistical methods as well as through spatial prediction techniques such as geographically weighted regression. Finally, the culmination of this review leads to a scientific and tested method with which the research questions of this research project were answered.

2.1 Property Values and Code Enforcement

Observations of most cities in the country reveal that there are areas with high property values and areas with low property values. Often times, areas with the lower values are the oldest in age, have the highest crime rate, have high occupancy turnover rates, and are very poorly maintained. The primary mechanism with which a municipality or other governing agency can attempt to address this is through enforcement of housing codes designed to maintain standard living conditions such as adequate lighting and ventilation, fire egress, heating and cooling, and proper roof maintenance (Meier 1983).

2.1.1. What is Code Enforcement?

Cities across the country employ code enforcement officers to ensure that both building and construction codes as well as municipal codes are properly met. These codes can be anything from requiring that a citizen obtain a building permit for their new patio cover to requiring that a citizen pull all of the weeds from their front yard or move the inoperable vehicle off the driveway and out of sight from neighbors. Should citizens choose to disobey the code and the code enforcement officer, the office has the authority to issue a monetary citation against the property with the violation as an incentive for the citizen to comply. If compliance is still not achieved, a recorded lien can be placed on the property preventing sale or transfer of ownership until fines are paid and compliance is met. These violations are reported in a manner similar to crime insofar as a citizen must report the violation or file a complaint with the city's code enforcement department. An officer must then respond to the report or complaint and enforce the necessary codes just as a police officer would enforce the necessary laws or penal codes. As the City of Victorville attempts to improve its image by cleaning up the residential properties through proactive code enforcement, special attention will be given to front yard aesthetics and

landscaping.

2.1.2. High Quality Landscaping Increases Property Value

There have been several studies conducted that show a positive correlation between landscaping and greenspace and its effect on property values. This is intuitive, especially for most homebuyers as homes that are the most visually appealing on the outside attract the most attention. Des Rosiers et al. (2002) looked at how tree cover, ground cover, lawn cover, and landscape structures increased property values for homes that had these amenities over homes that did not. They found that landscape features such as a hedge or flower arrangement increased homes values by up to 3.9% in some cases, while decorative structures such as patio covers increased property values by as much as 12.4%. The architecture of the building was even found to have a significant impact on the results as bungalow and cottage style homes with quality landscaping had the highest values over row houses with similar landscaping.

A study by Dombrow, Rodriguez, and Sirmans (2000) conducted in Baton Rouge, LA attempted to assess the value added by mature trees on a residential property. They used a multiple regression approach using home sale data from the area and found that regional assessors typically added around 2% to the total value of the home if it had mature trees on the property that were aesthetically appealing and provided shade and other benefits to the property. This coincides with a previous study done by Anderson and Cordell (1988) in Athens, GA that found that single-family homes with landscaping with trees subsequently saw an increase in sales prices by 3.5 to 4.5% over properties that did not. Both of these studies used a regression model and found statistically significant correlations between the sale price and the presence of landscaping and trees.

A study conducted by Luttik (2000) also looked at how natural landscapes can have

substantial effect on the value of a home. While the study of Des Rosiers et al. (2002) only surveyed 760 single-family homes in Quebec, Canada, Luttik (2000) surveyed almost 3,000 single-family homes over a large region of the Netherlands. Other than the number of homes in the survey, these studies were very similar. Both used a hedonic price model where they analyze the effect of the total price by specific property characteristics. In essence, they both determine the effect of landscaping, location, property size, and architecture style separately and then combine them into a total property value. While Des Rosiers et al. (2002) looked at property landscapes, Luttik (2000) looked at location adjacent to greenspace, view of greenspace, view of water features, and landscape diversity. They found that homes located directly on a lake had the highest increase in property value at roughly 12%, those that were bordering an area with a park, forest, or lake had increases of roughly 5% to 8%, and homes not located near greenspace adjacent to more urban areas with multi-story apartment buildings actually saw a decrease in property value by up to 7%. Unfortunately, the City of Victorville does not have any scenic water features; however, there are still greenspaces such as parks and golf courses that can have an impact on surrounding property values.

In addition to Luttik (2000), the study of Conway et al. (2008) attempted to model how property values are affected by greenspace in surrounding areas. Conway et al. (2008) expanded on the hedonic approach of Luttik (2000) to include a spatial lag model to account for the spatial autocorrelation determined to be present in the hedonic model. They discovered that even with spatial autocorrelation removed through the spatial lag model, there are positive impacts on property value from greenspace. However, the effect deteriorates with distance as greenspaces greater than 400 feet away from a property were shown to have negligible impacts on property value. This clearly shows that Tobler's First Law of Geography is in effect here, as objects in

close proximity to one another exert greater impact than objects further apart (Tobler 1970).

What these studies show is that there is a significant effect on property values from greenery and mature landscaping. Trees, water features, scenic views, and overall access to open space are shown to increase residential property values by roughly 3% to as much as 12% in some instances. In all of these cases, a regression model was employed in order to see how well these landscaping and open space characteristics affected the dependent variable of property value. Each one sought to determine which factors were most influential in order to make predictions on property values in future transactions. For this project, these studies act as a justification for the City of Victorville's current proactive code enforcement policies that are aimed at increasing property values by ensuring that residents maintain their property appearance with proper landscape maintenance.

2.1.3. Managing Residential Properties through Code Enforcement

For decades, cities of all sizes have attempted to keep residences within their jurisdictions in a good state of repair through code enforcement and other forms of criminal and legal processes. Many studies were done in the 1960s to analyze what needed to be done with regards to housing conditions. This came during a time of heavy housing reform that ultimately led to the creation of the Department of Housing and Urban Development (HUD) in 1965, which was charged with the regulation of housing in the United States. Housing acts such as the Fair Housing Act of 1968 and the Housing and Community Development Act of 1974 were also passed in an attempt to increase housing quality in the country. These acts also created Section 8 housing, which is a government program that helps renters offset their rental costs if they have a low enough income level to qualify (Teater 2011).

Two studies done during this time still have relevance in 2015. One such study published

in the Harvard Law Review by Carlton, Landfield, and Loken (1965) examined how municipal housing codes were enforced at the time and offered a different approach to more effective results. Another study from the Columbia Law Review by Gribetz and Grad (1966) looked at how code enforcement evolved over time, what the primary issues were at the time of the article, and what could be done to alleviate some code enforcement inefficiencies.

Both of these studies pointed out several issues with the overall code enforcement process, including the idea that full results are almost never achieved and that it relies on a criminal process that is often slow and inefficient. Carlton, Landfield, and Loken (1965) note for rental housing that code violations are often given to the property landlord and not to the tenant who is more often than not the party responsible for these violations. This leads to a conflict between the landlord and the tenant on who should ultimately take responsibility for the violation, which in turn disrupts the mitigation process. They also discuss the difficulty of enforcing housing codes across a municipality in an un-biased manner. This is largely due to the fact that many housing codes are broad and require much interpretation that gives way to unfair enforcement and potential corruption. The authors went on to describe the various "remedies" of code enforcement such as fines, liens, and a judicial process that are still present in today's code enforcement process. In addition, they describe how New York City also imposed rent control and withholding proceedings to enforce housing code violations.

These discussions were reinforced in Gribetz and Grad (1966) as they also discussed New York's rent control process, as well as the idea that the code enforcement process is slow and has many hurdles that must be cleared both physically and administratively. The researchers point out that the criminal code enforcement process does not work very well on major violations because there is often a financial barrier preventing an owner or tenant from correcting the issue.

This leads to a long and drawn out process of citations and inspections that ultimately increase financial burden due to fines and various other actions that require some form of monetary transaction that does not go directly to correction of the violation. They propose a stratified system of enforcement where infrequent violators are given more leniency versus a "hard-core" violator that must be taken to court.

Both of these articles give several insights into how housing code violations can be dealt with by the authority having jurisdiction. The laws and processes that were discussed here are still in effect today and cities across the nation, including the City of Victorville, still follow the process of citations, fines, liens, and ultimately criminal sanction to address the violations with cities. However, as the articles discuss, these tools and processes are often inefficient or slow which causes violations to linger and passes financial obligations for these processes on to property owners and other violators. This in turn causes a circular effect resulting in more required action by both the city and the property owners which leads to resentment and hostility on both accounts (Carlton 1965; Gribetz and Grad 1966).

2.1.4. The Effect of Code Enforcement on a City

Many have argued that code enforcement has either a positive or a negative impact on a city. On one hand, there is the idea that active code enforcement delays residences from falling into decay as they age, and code enforcement is essentially not needed in high income level neighborhoods. On the other, there is the notion that people living in run down residences do not have the disposable income to pay for routine maintenance of their homes. The result is a downward trend that is nearly impossible to escape. This section looks at both arguments.

The positive effects of code enforcement were studied by Ron Meier (1983) where he analyzed how the City of Pasadena, CA managed code compliance through an inspection

program that was required each time a residence changed occupancy. The city was divided using existing Census tracts and were then categorized by income level. He then analyzed the change in property value over time and compared it to how active the code compliance inspection program was in these tracts. The result was that upper-middle class homes were not affected by the code inspection program which saw a natural market value increase due to neighborhood affluence. The lower class neighborhoods that had the most active code inspection activity saw some increase in value from code compliance inspections, but it was not significant compared to middle class residences. This study found the most significant impact on property value came in these middle class neighborhoods where code compliance inspections uncovered minor violations that were easily fixed because either the cost of making these repairs was passed on in the selling price, or the landlord was able to take care of these issues before new tenants occupied the residence. These findings indicate that there is a need within Pasadena to allocate code enforcement resources to less affluent neighborhoods because the effect is more substantial.

The negative aspect of code enforcement was discussed in the work of Miller (1973), Ross (1996), and Burby et al. (2000). Each of these studies discussed how too much code enforcement can prove to be detrimental to code enforcement's goal of cleaner and safer neighborhoods. Miller (1973) discussed the economics of code enforcement with emphasis on the idea that some code enforcement is good, but there is a point where a breaking point is reached and property owners must simply "walk away" from the property, especially if they are renting. Miller (1973) looked at code enforcement in terms of five levels of enforcement with 1 being the lowest and 5 being the highest. He notes that at level 1, there is no financial burden placed on the owner but there is no oversight on how that property is maintained. At level 2 and 3, there is some financial burden, but the property is maintained because of pressure from

inspectors or officers. At level 4, the property owner begins to take a loss on the property because the financial burden of code enforcement requirements, property taxes, and other costs exceeds what they earn in rent, but they continue to hold the property. At level 5, the financial burden is too great and the owner walks away and abandons the property or sells it. At the neighborhood scale, a code enforcement level of 5 would cause more problems than it would fix because there would be a high level of abandoned properties that would be subject to vandalism and blight.

In addition, the work of Ross (1996) and Burby et al. (2000) further emphasize that too much code enforcement leads to urban decline. Both of these studies give examples of how too much government oversight or too much discretion on the part of the inspectors ultimately places too much of a burden on developers and property owners to make it feasible for them to continue building or owning property. Furthermore, they both discuss that code enforcement must be tailored to each neighborhood based on social and economic factors in order for it to succeed in improving neighborhood quality while at the same time encouraging residents to remain. Ross (1996) goes on to point out that overbearing code enforcement leads to property abandonment, just as in Miller (1973), but he also notes that abandonment is contagious and can cause low income neighborhoods that are otherwise stable to change into slums that are ridden with crime and drug abuse.

2.2 Crime Prediction

Crime enforcement and code enforcement share several similar characteristics. Both require an officer to respond to a citizen's call or complaint, both require a location to respond to, and it is advantageous to be able to predict where these instances will happen. For this reason, this section takes a look at the work that has been done in geographic information

systems (GIS) to predict the occurrence of crime in order to draw similarities and make assumptions as to how various techniques can be applied to code enforcement prediction.

2.2.1. Traditional GIS Based Approaches

There are various methods in GIS that can be used to analyze points and visualize patterns. Murray et al. (2001) outlines several different spatial analysis techniques for understanding patterns in crime data for Brisbane, Australia. These techniques include exploratory analysis, cartographic display, and optimization-based clustering. These techniques are largely visual in nature and require that the analyst has knowledge of the data being displayed. Furthermore, these techniques make prediction difficult because underlying relationships may not be displayed in the output maps. Murray et al. (2001) goes on to discuss the various spatial statistical analysis techniques that have more predictive capabilities. Global techniques, such as Moran's I and box maps, depict location based relationships across an entire study area. This can potentially give the same weight to features far from a subject point as it does to a nearby feature. Localized techniques such as Moran's scatter plot and local indicators of spatial autocorrelation (LISA) place more emphasis on nearby features. These techniques show spatial clustering and statistical significance among the data.

2.2.1.1. Hot Spot Mapping and Analysis

The most common form of crime analysis is hot spot mapping. This method can be employed in many ways. Clustering approaches using hierarchical and partitioning techniques allows analysts to see crime clustering in small geographic areas and group them based on pre-specified criteria. The major disadvantage to this technique is that it is difficult to distinguish the number of significant clusters in the data (Grubesic and Murray 2001). Studies have looked to improve the traditional hot spot analysis.

The work of Liu and Brown (2003) used a newly developed density model that analyzes transitional density as a means of hot spot detection and compared it to standard hot spot techniques. Findings indicated that the transitional model outperformed the hot spot models in all but one comparison. They were essentially able to improve upon hot spot detection by incorporating density estimates over time and space and by isolating features within the data that had the most explanatory power. This was also done by Xue and Brown (2004) where they incorporated the assumption of criminal preference in burglary data and found that this significantly improved on the hot spot detection techniques. These studies clearly show that simple clustering and hot spot detection do not always depict all spatial relationships in data and that there are better techniques available.

2.2.1.2. Kernel Density Estimation

One technique that improves upon hot spot mapping is kernel density estimation. In the writing of Chainey et al. (2008), they used kernel density estimation in conjunction with standard deviational ellipses, thematic maps and grid analysis to test the accuracy of hot spot analysis as a spatial predictor of crime. The kernel density estimation out-performed all other techniques based on predictive accuracy index. Nakaya and Yano (2010) take kernel density estimation to a new level in crime analysis in Japan. They employed a new technique of 3D visualization in a space-time cube where the variable of time is added to the typical kernel density estimation. What this does is show temporary association between two known crime hot spots that may only exist for a short period of time.

Crime prediction using kernel density estimation can also be employed using recent spatial data from social media, such as Twitter. This media captures the geographic location of an individual's "tweet" along with the text within the tweet itself. A study in this phenomenon

was conducted by Gerber (2014) in Chicago, IL to predict the occurrence of crime. Gerber (2014) created a training model using one month's worth of "tweets" specifically posted for crimes using kernel density estimation and found that the standard kernel density estimation was significantly improved for 19 out of 25 crime types studied. He did caution that it is unclear as to why the Twitter data enhanced the kernel density estimation, but this study still demonstrates how kernel density can be used to predict crime using non-traditional data sources.

2.2.2. Regression Modeling Techniques

Crime has also been predicted using various regression methods. Bayesian regression (Wheeler and Waller 2009), Tobit regression (Osgood, Finken, and McMorris 2002), and logistic regression (Antolos et al. 2013) have all been used in the past. Tobit regression is intended for data that is "censored," meaning that there is an obstacle or a value limitation. Tobit regression was actually shown to improve standard modeling techniques such as Ordinary Least Squares (OLS) by eliminating unnecessary components of the model. Bayesian regression is used as a hierarchical approach to correct increased coefficient variance created when using geographically weighted regression (GWR). Both of these methods have their merits, but the logistic regression technique is what was employed in this research project because logistic regression is used for binary dependent/response variables (Antolos et al. 2013).

2.3 Logistic Regression Techniques

Logistic regression modeling techniques can be employed in any field where the final outcome is dichotomous or binary. Examples include yes or no, present or not present, dead or alive, and 0 or 1. This section looks at three studies where logistic regression was utilized to make a prediction without employing any kind of geographical weighting factors. These studies employ similar methodology in the fact that they use a stepwise regression model using portions

of a known dataset to train their model while the other portion of the known dataset is used to validate the predictions made by the model.

The first study performed by Perestrello de Vasconcelos et al. (2001) attempted to predict the probabilities of wildfire ignition in Portugal. They sought to compare the outcomes of a logistic regression model and a neutral networks model. A stepwise logistic regression model was created using several environmental variables such as topography and distance to man-made features to find a probability of ignition value. They found that although the logistic regression model proved to be a strong predictive tool, the neutral networks method made more significant predictions and was more robust. This research project loosely followed the logistic regression model techniques employed here, specifically using the logistic regression model to determine which variables provide the most significant predictive capabilities. However, because this study was more of a comparison between two different methodologies, this research project did not follow it directly because it does not factor in geographical variability in the data.

The work of Ozdemir (2011) employed similar methodology as Perestrello de Vasconcelos et al. (2001) with regard to using a stepwise logistic regression. However, Ozdemir (2011) uses the technique as the primary function for predicting groundwater spring potential in Turkey. Here, he uses variables such as land use, lithology, slope aspect, and elevation in raster data to create a spring probability raster which showed the areas that are likely to support a spring and which areas are not. The data training points in this research consisted of a set of known control pixels within the landscape raster that contained a spring and an equal number of pixels from the areas in the raster that are not known to contain a spring that were chosen by a random selection tool within ArcGIS. The logistic regression model created here had an overall predicted accuracy of approximately 95%, which indicates that it is an extremely strong model.

Unfortunately, the methodology used in this study cannot be used in this research project because it used raster data to predict if a spring could or could not be present. However, as with Perestrello de Vasconcelos et al. (2001), this research project utilized the model validation principles of checking predicted outcomes with known control observations.

A similar study conducted by Kundu et al. (2013) also used logistic regression techniques to predict landslide occurrence using raster data. The primary difference from the two previously discussed works involving logistic regression was that this study used 31 explanatory variables in the model. Simplicity becomes the issue here. In the study, the researchers used the forward stepwise method with their logistic regression model which began by calculating the model with no explanatory variables, followed by introducing only statistically significant variables one at a time until all of the significant variables were in the model. Variables that were not statistically significant at the 0.05 level were thrown out. Kundu et al. (2013) kept only seventeen of the thirty one initial explanatory variables, which is nearly half. Even though Kundu et al. (2013) employed the SPSS software program, their methods were not utilized for the purposes of this research project because they used raster data, there was a very large number of explanatory variables, and the sampling method was similar to Ozdemir (2011). However, using the forward stepwise method of logistic regression was considered when determining which explanatory variables were to be used within the context of this research project.

2.4 Geographically Weighted Regression Techniques

Spatial location plays an important role in regression modeling. One regression technique that factors spatial location is geographically weighted regression (GWR). Atkinson et al. (2003) employed GWR to model riverbank erosion in Wales using various geomorphological variables such as streamflow, material flow, meander, history, and vegetation. They used a

univariate method of GWR, and explored a logistic form of GWR to fit their model to each explanatory variable rather than fitting many variables to one model as in multivariate regression. They did note that the Gaussian model was not used in this analysis because it only explains spatial autocorrelation between variables. Their primary goal in this study was to determine relationships rather than predicting outcomes, therefore, this methodology was not utilized for the purposes of this research project. However, Atkinson et al. (2003) was useful in understanding how GWR can be used to determine spatial relationships between explanatory variables and that GWR can be used for other academic purposes beyond prediction.

The predictive capability of GWR was explored in Erener, Sebnem, and Düzgün (2010) where they compared logistic regression and neutral network techniques to geographically weighted regression techniques to show which method was the strongest predictive tool. Their primary purpose was to improve upon the non-spatial methods of logistic and neutral network models by incorporating spatial factors. Their study involved prediction of landslide susceptibility in Norway and included variables such as precipitation, slope, aspect, geology, and others. Many of these variables were raster data; therefore, the methodology of Erener, Sebnem, and Düzgün (2010) was not used in this research project. However, their findings are what was of interest. Even though they found that the logistic regression technique was reliable and provided strong predictive capabilities, the GWR model proved to be stronger. Their study yielded pseudo R squared values between 0.14 and 0.19 for the logistic regression model, and 0.54 for the GWR model. This was a profound improvement. Erener, Sebnem, and Düzgün (2010) also note that the predictive accuracy of the GWR method was greater than that of logistic regression. This study is a great example of how GWR improves upon traditional non-spatial techniques, which is the reason that this research project is utilizing geographically weighted

logistic regression to predict if a single-family home has or does not have a code enforcement violation.

Another study that compared logistic regression and GWR was that of Saefuddin, Setiabudi, and Fitrianto (2012). Here, they examined the differences between a global logistic regression model and a local geographically weighted logistic regression model in predicting the poverty level of regions in Indonesia based on the Human Development Index. The global logistic regression method was performed first, followed by the local geographically weighted logistic model. For the global logistic model, the researchers note that the odds ratio is a better way to interpret the results of the model because it is easier to interpret than the model coefficients. This was considered in this research project when interpreting the results of the logistic regression output for the areas that have already been proactively surveyed by code enforcement. For the local geographically weighted logistic regression model, Saefuddin, Setiabudi, and Fitrianto (2012) note that the weighting function and bandwidth distance selections are highly important in the GWR model because these selections determine how the spatial location of other instances in the data affect the instance being analyzed. They note that the Gaussian weighing function was problematic because of the difficulty of assigning weights to all data in the study area. They recommended using the bi-square function instead because of its ability to remove points in the dataset where the distance between the subject point and its weighting point is greater than the selected bandwidth distance. Essentially, the bi-square method calculates the regression using only data points that are in the specified bandwidth distance. This was considered for the purposes of this research project when preparing the geographically weighting logistic regression model in the GWR4 software. Finally, Saefuddin, Setiabudi, and Fitrianto (2012) found that GWR was a superior model to logistic regression

because it was able to fit data points more accurately and was the best method for their analysis of poverty.

Sections 2.3 and 2.4 are good examples of how logistic regression and geographically weighted regression can be combined to make a stronger predictive model. This research project followed suit with the aforementioned studies and utilized multiple aspects of their methodologies. Chapter 3 discusses four additional works where logistic regression and GWR were used to make predictions for spatially occurring phenomena and how the methodology of these studies was utilized to compose the methodology of this research project. It also discusses the data needs and variable justification for the model that was created to predict code enforcement violations.

Chapter 3 Data and Methods

This study was undertaken to determine if it is possible predict if a single-family residence (SFR) in the City of Victorville (herein referred to as the City) is likely or not likely to have a housing code violation in an effort to make the code enforcement process more efficient. Similar to crime prediction, this study looks at both spatial and non-spatial factors that have the potential to predict this outcome. Since the prediction is a binary outcome (yes or no), a non-spatial logistic regression model was used to determine the key explanatory variables that have the greatest influence on this outcome. Spatial factors were subsequently considered by employing geographically weighted logistic regression using the key explanatory variables from the non-spatial logistic model. The result was a model that was used to determine which of the 29,000 plus single-family residences in the City are likely to have a housing code violation.

This chapter explores the methodology employed to reach a prediction of either yes, a housing code violation is likely, or no, a violation is not likely for every SFR in the City. The methodology was based in part on studies done by Lee and Sambath (2005), Wu and Zhang (2013), Martinez-Fernandez, Chuvieco, and Koutsias (2013), and Rodrigues, de la Riva, and Fotheringham (2014) where both non-spatial and spatial logistic regression techniques were utilized to determine the binary outcome of environmental factors. In addition, it explains how the independent variables were determined, the process of the non-spatial logistic regression technique and how its primary assumptions were met, and the process of determining the spatial variability in this study.

Chapter 3 is composed of three sections. The first section describes the non-spatial and spatial logistic regression techniques and how the methodology was selected for this study. The second section discusses the data that was needed for this study and how it was acquired, and it

discusses justification for each of the independent variables. Finally, the third section discusses the procedures that were followed to reach a predicted outcome for each SFR in the study area.

3.1 Research Design

An important aspect of this research project was determining which regression technique to use in order to predict if a SFR has a housing code violation. Many of the variables that were selected were categorical and non-continuous, eliminating the ability to do the more widely utilized ordinary least squares (OLS) and geographically weighted regression (GWR) techniques because with those techniques all variables must be continuous and they must be on interval or ratio scales. Furthermore, the traditional linear regression technique of statistical prediction could not be employed because the outcome is binary. Therefore, the method chosen for this study was binary logistic regression because it capably handles the binary outcome, nominal and non-continuous variables, and binary variables.

3.1.1. Non-spatial Logistic Regression Technique

The binary logistic regression (BLR) technique attempts to fit a set of data along a line of best fit, similar to a traditional linear regression. However, this line of best fit does not follow a straight line, but instead follows a logarithmic "S" shaped curve because the binary outcome of yes/no or 1/0 causes points to be concentrated around two locations on the curve (see Figure 3). Furthermore, the BLR model assumes the independent variables are non-linear functions of each other, meaning that multicollinearity should be checked prior to beginning the regression analysis, which is similar to the OLS model. BLR also assumes the actual dependent variable and the independent variables do not share a linear relationship, but rather a linear relationship between the logit of the dependent and independent variables. Other assumptions include the fact that the dependent variable does not have to be normally distributed, the independent

variables do not have to be interval or ratio scale, and there needs to be a large sample size (Anderson, 1982). The prediction equation of the BLR model uses the odds ratio and a constant to compute the probability that a dependent variable value will fall between either 1 or 0, or yes or no. The cut off value is typically set at 0.5, meaning that if the probability is greater than 0.5, the value is assumed to be 1/yes and if it falls below 0.5 it is assumed to be 0/no.

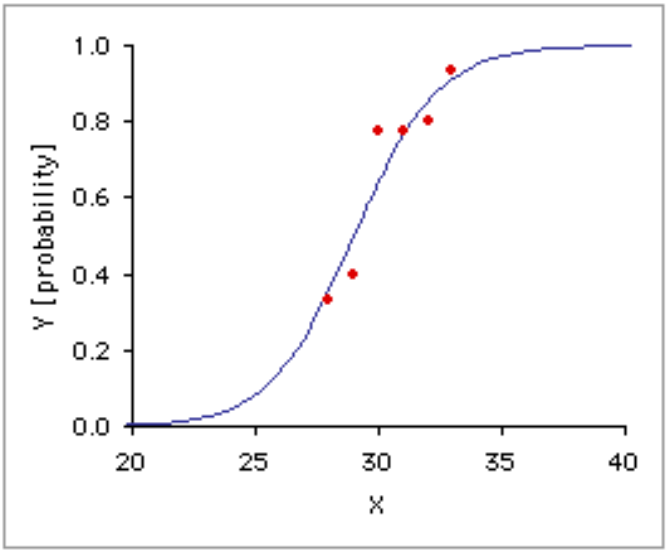


Figure 3: A simple logistic regression curve

The binary logistic regression prediction equation is given in Equation 1 below, where $\log(p_i / 1- p_i)$ is the prediction percentage, α is the constant, β is the variable coefficient, and X is the variable value.

$$\log\left(\frac{p_i}{1-p_i}\right) = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} \qquad (1)$$

Lee and Sambath (2005) used a logistic regression model to study landslide susceptibility. Here they used the logistic regression modeling tools available in IBM's SPSS statistics software to train a model using known landslide data and multiple independent variables to then predict which regions in their greater study area were most susceptible to

landslide. The initial portion of this study largely followed the methodology of Lee and Sambath (2005) where the SPSS program was used to train a logistic regression model to be used in predicting the likelihood of a housing code violation for 29,000 homes using a sample of 2,300 homes within the city.

3.1.2. Geographically Weighted Logistic Regression Technique

The geographically weighted logistic regression (GWLR) technique introduces a spatial aspect to the binary logistic regression model by including the X and Y coordinate points to the logistic regression equation. Essentially, a logistic regression is created for each instance in a spatial dataset based on a selection of surrounding instances. A distance band, or kernel, must be specified to determine how much influence each occurrence exerts on the others. This kernel determines the number of surrounding data points that get factored into the regression equation of the data point being regressed. This measure is crucial in the GWLR technique and must be calculated carefully because it determines the degree of spatial influence in the GWLR equation.

The GWLR equation (2) is given below, where $\log(p_i / 1- p_i)$ is the prediction percentage, α is the constant, β is the variable coefficient, (u_i, v_i) are the coordinates of the variable, and X is the variable value. The surrounding points within the kernel distance are weighted against the data point being calculated so that points close to the subject point exert more influence than points that are further away. The (u_i, v_i) value in the equation is essentially a second coefficient value that expresses the geogaphic influence exerted on the variable value once the degree of influence from the surrounding data points within the kernel is determined. The value of (u_i, v_i) is the result of the weighting function applied to the location of the data point being regressed.

$$\text{logit}(p_i) = \log\left(\frac{p_i}{1-p_i}\right) = \beta_0(u_i, v_i) + \beta_1(u_i, v_i)x_1 \quad (2)$$

The studies of Wu and Zhang (2013), Martinez-Fernandez, Chuvieco, and Koutsias (2013), and Rodrigues, de la Riva, and Fotheringham (2014) all utilize the GWLR technique to model various environmental phenomena. Wu and Zhang (2013) began by using the non-spatial logistic regression method first to train their model, then compared that to the results of the GWLR method and found that GWLR created a more effective prediction model because it accounts for more of the spatial heterogeneity of the occurrences. Martinez-Fernandez, Chuvieco, and Koutsias (2013) followed similar methodology to model the occurrence of wildfire. They first modeled the fire occurrence using linear and logistic regression, as well as ordinary least squares before utilizing geographically weighted regression to measure the effect of spatial relationships in their data. In this case, traditional GWR was used instead of GWLR because their dependent variable was not binary. However, they did use GWR3 software to compute the spatial aspect of this model because it was more robust than the tools available in the ArcGIS software available at the time. This was a contributing factor in the selection of the GWR4 software used for this research project because it allowed the analysis to go beyond the ArcGIS platform where tools for more defined GWR analyses, such as geographically weighted logistic regression, are not yet available. Finally, Rodrigues, de la Riva, and Fotheringham (2014) followed the same methodology as the former, but instead used a random sample of wildfire occurrences to train both their logistic regression and GWLR models.

This research project followed the techniques of these four independent studies. A non-spatial logistic regression model was created using a sample of approximately 2,300 single-family homes from three separate neighborhoods within the City. This model was then used to

determine which independent variables were statistically significant, which is a key assumption within logistic regression. The analysis then moved to GWLR within the GWR4 software obtained from Arizona State University to measure the spatial variability within this model and to make housing code violation predictions for the entire City's single-family residences. This process is outlined in more detail in Section 3.3: Procedures and Analysis.

3.2 Data Requirements & Data Sources

Data from several sources were needed for this research project. The San Bernardino County Assessor's Office and the City of Victorville were the primary sources of the required data, as well as the California Department of Alcoholic Beverage Control. Most of this data was free to download over the Internet; however, some data from the Assessor had to be purchased and data from the City of Victorville had to be manually input from paper reports. The cut-off date for data collection was July 11, 2015, which is the day both the City's code enforcement data was collected and the date that Assessor data was downloaded. Data beyond this point in time was not considered even though analysis of the data took place two months later.

The San Bernardino County Assessor's Office provided much of the data in the form of an Assessor's parcel feature class that is available to download for free on the County's website. This data contained an Assessor's parcel number (APN) that acted as the primary database key to match a property's attributes from other sources, and the situ address that was used to join properties from the proactive survey. The Assessor's Office also provided street centerlines shapefile for free, which was used to compute two independent variables. Other Assessor data was collected through Mimi Song Company of Ontario, California using their access to First American Title Company's MetroScan program which links to the Assessor's live database.

The City of Victorville provided the proactive code enforcement data, previous code

enforcement data, as well as rental property business license data used to determine the type of occupancy of the property. California's Department of Alcoholic Beverage Control provided location of liquor stores that was used in conjunction with street centerlines to compute distance based variables.

3.2.1. Dependent Variable

The dependent variable in this study was the City of Victorville's proactive code enforcement survey. Through the City's current proactive survey, code enforcement officers must visit each property within a target neighborhood and note the presence of any violation seen. These violations can range from unmaintained landscaping, presence of in-operable vehicles, illegally parked vehicles, dilapidated building conditions, or unpermitted structures. Officers then give the resident a written notice of the violation(s), retaining a copy of this notice for the administrative records. At the time of data collection for this study, code enforcement staff had surveyed three distinct neighborhoods outlined in Section 3.3.1 later.

In order to convert these reports into spatial data, each address of the cited properties had to be logged into a spreadsheet, parsed, and recombined into a format of *address number* and *street name* to successfully join it to the parcel feature class. This was done for all three neighborhoods in the same manner. Properties within these neighborhoods were selected based on the streets noted in the survey area. Selected properties in the proactive survey were given an attribute value of "YES" while properties that did not appear in the proactive survey were given an attribute value of "NO," thus forming the binary dependent variable.

3.2.2. Independent Variables

The independent variables for the model were created using several different processes. Some were obtained directly from the data source, others were obtained through simple

calculations using other variables, and others were obtained using network analysis. Table 1 lists these variables.

Table 1: Independent variable list

Independent Variable
Property Size (Square Feet)
Total Property Value (Dollars)
Value Per Building Square Foot (Dollars)
Floor Area Ratio (Building Size \ Property Size)
Corporation Ownership (Yes/No)
Occupancy Type (Renter/Owner)
Number of Building Stories (1/2)
Structure Age (Years)
Length of Ownership (Days)
Tax Default Status (Yes/No)
Previous Code Case Present (Yes/No)
Number of Cases (2005 to Present)
Days Since Previous Violation (Days)
Number of Neighbor Cases
Distance to Liquor Store (Cartesian) (Feet)
Distance to Liquor Store (Network) (Feet)

3.2.2.1. County Assessor Data

Variables collected directly from the San Bernardino County Assessor's office were the *size of the structure*, the *corporation ownership (yes/no)*, the *tax default status*, and the *number of building stories*. Structure size, ownership type, and building stories were obtained by downloading the data through the MetroScan program for all Victorville parcel numbers. The tax status variable was obtained by purchasing the San Bernardino County Tax Collector's TR345 report, which is only available on CD. The parcels contained in this report were joined to the parcels feature class based on APN and exported to create a separate feature class.

Other variables were derived by computing County data with other County, City, or geometric data. The *total property value* variable was computed by adding the land value and

structure value attributes provided in the parcels feature class using field calculator. The *value per building square foot* was computed in *Field Calculator* by dividing the total property value by the square footage of the structure. It does not include the square footage of the garage. The *property size* was calculated using the *Calculate Geometry* option in ArcMap using the State Plane Coordinate System CA Zone V NAD 1983 Datum. The *floor area ratio* was computed by adding the structure square footage and the garage square footage together, then dividing by the property square footage. *Structure age* was calculated by subtracting the "year built" field from the Assessor data by the current year (2015), leaving the age of the structure in years.

3.2.2.2. City Data

City data are all available through a Request for Public Records document through the City's Records Department at no cost. Data was extracted through the City's permit and licensing software called "Tidemark." To do this, Tidemark had to be queried using the *MS Query* function in Microsoft Excel. Data for rental homes was extracted by using the business license category code for rental units and legacy data for previous code enforcement cases were extracted for the entire database, which dated back to early 2005.

Rental licenses were joined to the county parcels feature class based on APN and exported as a separate feature class before being spatially joined to the feature class containing the independent variables in this analysis. Parcels that were successfully spatially joined were given a value of "renter" in the new *occupancy* field, and the remaining parcels were given a value of "owner." It is worth noting here that the rental licenses used in this variable were only the rental homes known to the City at the time of this analysis. Unfortunately, this database does not account for all of the rental homes in the City, but it does account for most. The actual number of rentals in the City has not been determined.

The code enforcement data was used to create the *previous code case present, days since previous violation, code case count,* and *number of neighbor cases* variables. Previous code case present was simply a yes or no binary variable indicating that the property had at least one housing code violation in the last ten years. If a case was successfully joined to a parcel, it was given a value of "yes," and a value of "no" was given to the remaining parcels. Days since previous violation was calculated in Excel by sorting date values chronologically and removing duplicate APN rows for older cases on the same parcel. This left the most recent case, and the number of days between the case date and July 11, 2015 was calculated before being joined to the rest of the variables. Code case count was also created in Excel by using the *Consolidate* tool, which essentially counted the number of cases associated with each APN. Number of neighbor cases was created by using a buffer selection of parcels containing a code case and appending that count as a new field in the feature class containing the study variables.

3.2.2.3. Other Data Sources

The California Department of Alcoholic Beverage Control (ABC) provided the location of commercial establishments that sell alcohol for off-site consumption. All locations within Victorville were selected, as well as those in the surrounding cities of Hesperia, Apple Valley, and Adelanto in order to ensure that the closest liquor store was chosen for each SFR in the city. A simple street network was constructed using *Network Analyst* and the street centerlines feature class provided by the County of San Bernardino. The network was then solved to calculate the travel distance in feet from each SFR to the nearest liquor store location. In addition, the *Near* tool in ArcGIS was used to calculate the Cartesian distance from each SFR to the nearest liquor store location. This created the *network distance* and the *near distance* variables that were then appended to the variables feature class.

3.2.3. Variable Justification

Variables in this analysis were chosen for a number of reasons. Some pertained to the building itself, some pertained to the occupants, and others to the neighborhood of the SFR. Structure age and total property value were natural choices because in most cities, buildings constructed under less thorough building code standards are most likely to become dilapidated and lose their value. The property size was selected to determine if neighborhoods that did not have a lot of space between buildings had a higher likelihood of housing code violations while neighborhoods with ample space between buildings were less likely. The value per square foot and floor area ratio variables were created to normalize the total value variable and the property size variable, respectively.

With regards to the SFR occupants, the property ownership type, occupancy type, length of ownership, and tax default status variables were selected. Property ownership type was chosen to provide insight into whether or not the entity who owned the building played a role in how well it was maintained. The thought here was that if a building was owned by married persons or single persons who lived at the residence, it would be better maintained than a residence owned by a corporation from a different region. The occupancy type variable was selected to determine if there was a difference in the level of building care between renters, who often have high turnover rates, and the owners of the SFR, who often stay for years and have a certain level of pride in their home, generally. Length of ownership sought to expand upon this idea that ownership stability led to a high level of building care. Tax default status was chosen because intuitively if a resident does not have money to pay their property taxes, they likely do not have money to pay for maintenance of their residence. Code enforcement related variables were also added to account for occupants who have a tendency of property neglect or multiple offenses.

Neighborhood variables were used to determine if there was any effect on a property based on the quality of the neighborhood it is situated in. The number of nearby code enforcement cases was used to account for Tobler's First Law that if a property that was surrounded by code enforcement prone properties, it would be likely to have a code violation as well. This was obtained by using the *Average Nearest Neighbor* tool in ArcGIS to determine the average distance between SFRs' parcel centroids in each study area, which was then used as the selection distance for each residence to select the number of cases within that distance. This effectively accounted for adjacent SFRs to a subject SFR, even if they were separated by a street or alley on any side.

The distance measures to the nearest liquor store were created to introduce a crime element as it is noted by many studies that liquor stores create crime hot spots that can have a spill-over effect to nearby properties (Block and Block, 1995; Speer et al. 1998; Britt et al. 2005; and Toomey et al. 2012). This creates a strain on property values, which is what Gibbons (2004) and Linden and Rockoff (2008) show in their analysis of crime rates and property values. As discussed in Chapter 2 of this document, lower property values increase the likelihood of housing code violations.

3.3 Procedures & Analysis

The following discussion outlines the steps that made up this research project. Three individual neighborhoods were analyzed first, followed by five random samples of all three, and concluded with an analysis of the entire dataset, creating nine distinct analysis areas. A non-spatial logistic regression was conducted for all nine, followed by a geographically weighted logistic regression analysis. Areas 1, 2, and 3 were analyzed individually because they were three distinct datasets collected by code enforcement staff at different times. They also consisted

of different sample sizes, 581 homes, 316 homes, and 1,301 homes respectively. Sample size is an important factor in logistic regression, which was tested here. Five random samples of Areas 1, 2, and 3 were analyzed to determine if single-family homes were consistent across the three areas or if their neighborhoods played an important role in their variability. Finally, all homes in the three study areas were analyzed to compare how the model performed to the other, smaller, samples of data.

3.3.1. Study Areas

The study area for this research project was the City of Victorville in the Mojave Desert of Southern California (Figure 1). Within the City, three distinct neighborhoods were analyzed. Area 1 (shown in Figure 4) on the north end of the City consisted of the Cypress Point housing tract which was approximately 60 percent built out at the time of this study. The neighborhood consisted of moderate grade construction homes roughly two to eight years in age on parcels ranging in size from 6,000 to 10,000 square feet on average. The neighborhood has a homeowners association, so residents have a fair amount of resident oversight when it comes to the aesthetics of the neighborhood. Area 2 (shown in Appendix A) on the east end of the City consisted of two separate neighborhoods that were in close proximity to each other. One neighborhood had low-grade construction homes that were approximately ten to fifteen years old, while the other neighborhood had moderate to upper end construction homes that ranged in age from brand new to fifteen years old. Area 3 (shown in Appendix A) was on the west end of the City and consisted of low to moderate construction grade homes approximately ten to fifteen years old. Each area had unique characteristics and residents of varying economic status.

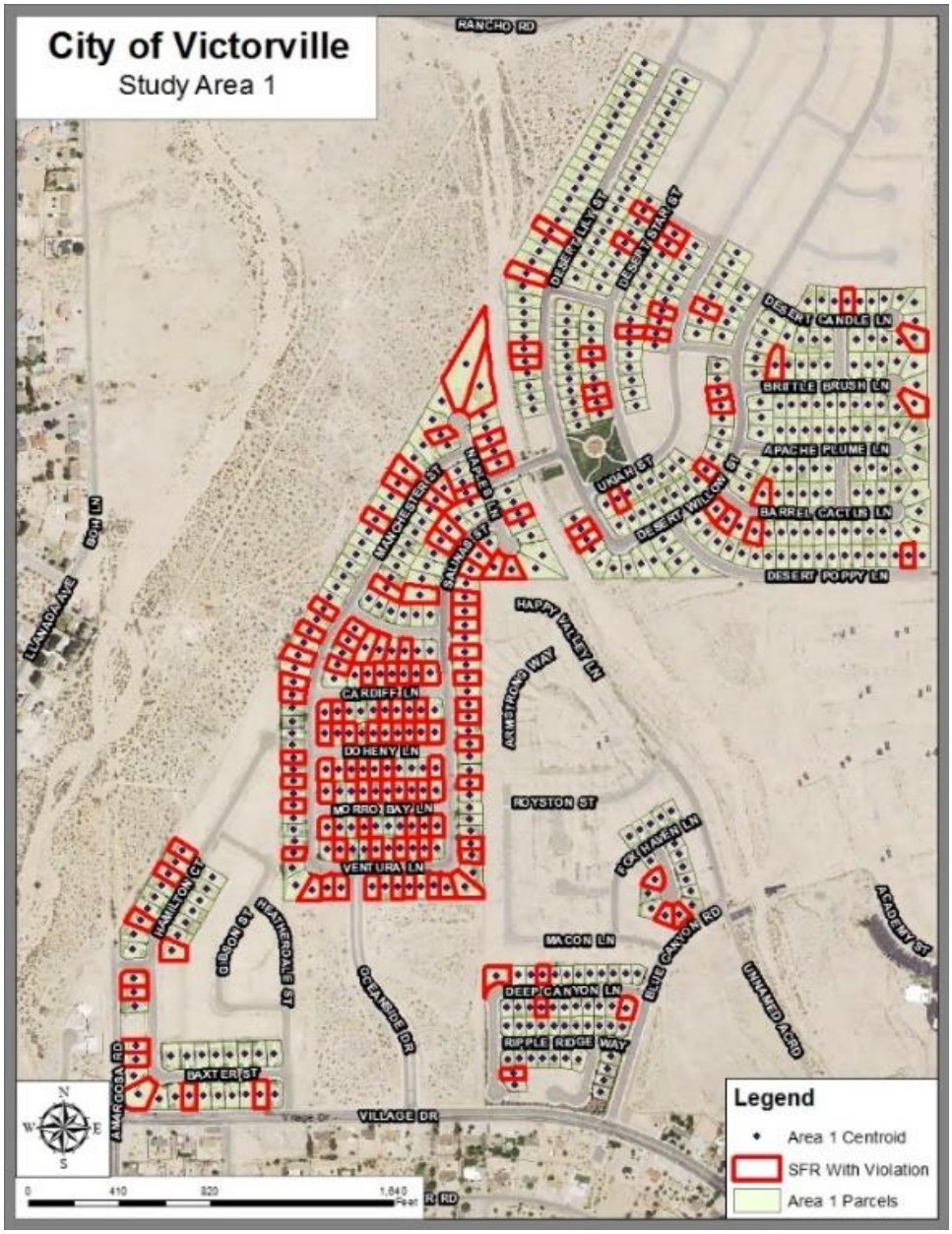


Figure 4: Area 1 neighborhood and observed violations

3.3.2. Individual Analysis of the Three Study Areas

Variables were first examined for multicollinearity in order to not violate a key assumption of logistic regression that collinear variables are not present in the model. This was done by performing a linear regression using only the independent variables. In IBM's SPSS 22.0 software, the linear regression tool was used for this purpose. The first independent

variable was placed in the dependent variable dialog box with the remainder of the independent variables placed in the independent variable dialog box. The test was iterated by removing the most recent independent variable from the dependent variable dialog box and swapping it with one from the independent variable dialog box until all independent variables were tested. The test produced *variance inflation factor*, or VIF, values explaining how each of the variables was related to the others. A VIF above 3.0 indicated that there was some degree of multicollinearity, and a value above 5.0 indicated that multicollinearity is present. For this research project, variables with a VIF above 3.0 were removed from the analysis.

A non-spatial logistic regression was then performed on each study area individually. Area 3 was done first because it had the greatest number of homes surveyed, and since one of the assumptions of logistic regressions is a large sample size, this was a logical starting point. The dependent variable and independent variables were uploaded to the SPSS software program and the logistic regression was performed using the "enter" method, which utilizes all of the independent variables in the model. After the initial iteration of the model, variables that were not statistically significant predictors at the 0.05 level according to SPSS were removed before running the model a second time. Statistical significance is one of the key assumptions of the logistic regression model and is used to determine the "goodness-of-fit" of the model. Chapter 4 explains these outputs and what they mean for the model.

Area 1 and Area 2 were then analyzed in the same manner as Area 3 where multicollinearity was first tested using simple linear regression, followed by the "enter" method for logistic regression.

3.3.3. Combined Area Analysis

The three areas were then combined into a single dataset and five random samples of 500

homes were taken using the random sample generation tool within SPSS. For each random sample, the same procedure was followed in the individual area analysis; non-significant variables were removed and multicollinearity was addressed.

A logistic regression was then performed on the entire dataset to compare how it performed to the individual analyses of Areas 1, 2, and 3, as well as on the five random samples. Again, this procedure followed the same steps as the previous analyses where multicollinearity was addressed using a linear regression of the variables, followed by the logistic regression analysis.

3.3.4. Geographically Weighted Logistic Regression Analysis

The role of geography was tested in this analysis by using the GWR4 software program to perform a geographically weighted logistic regression (GWLR) analysis. This was performed in the same manner as the non-spatial logistic regression analysis, where Areas 1, 2, and 3 were analyzed separately, followed by the five random samples of the data and concluding with the dataset as a whole. However, since GWR4 cannot handle nominal scale data, dummy variables had to be created for the dependent variable as well as for each categorical independent variable that passed the multicollinearity tests. Table 2 gives a breakdown of these dummy variables.

Table 2: Dummy variables for GWR4

SPSS Variable Value	GWR4 Variable Value
Occupancy Type	
Owner	1
Renter	0
Default Status	
Yes	1
No	0
Previous Code Case Present	
Yes	1
No	0
Corporation Ownership	
Yes	1
No	0
Proactive (Dependent)	
Yes	1
No	0

Within GWR4, the model was set to *Logistic (binary)*, the latitude and longitude coordinates for each single-family home's parcel centroid were input and the *spherical* option was selected. Next, the independent variables that remained at the end of the non-spatial logistic regression model for each area were input as local variables. The kernel type was set to *Adaptive Gaussian* using the nearest neighbor method and the bandwidth selection method was set to the *golden selection method* where the GWR4 software determines the optimum bandwidth measure for each dataset. This was done because the *incremental spatial autocorrelation* and *Moran's I* tools in ArcGIS did not produce any peak values for any of the datasets, meaning there is no distance where spatial relationships are most pronounced, according to these ArcGIS tools. Finally, the selection criteria was set to the *Akaike Information Criterion* (*AIC*) because the GWR4 manual states that this is the most suitable method when using a Gaussian kernel type. These parameters were set for each iteration of the GWLR analysis over the study areas.

Once each area was run through the GWR4 software, the output table containing variable coefficients and predictions was analyzed for accuracy to compare against the non-spatial logistic regression model. This was done by comparing the *y* column in the output table, which contained the actual observed dependent variable value for each case, with the *yhat* column which contained the predicted probability value for each case as calculated by GWR4. *Yhat* values below 0.5 were re-valued at 0 and values above 0.5 were re-valued at 1. If *y* and *yhat* matched, a new value was calculated at 0 meaning that the model correctly predicted the dependent variable. If the *y* and *yhat* did not match, a new value of 1 was calculated indicating that the model did not correctly predict the dependent variable. The total number of incorrectly predicted cases was subtracted from the total number of observed cases, and the value was then divided by the total number of cases to arrive at a prediction accuracy percentage number. The prediction accuracy percentage number was then compared to the same number calculated in the logistic regression model found in SPSS.

3.3.5. Model Validation and Making Predictions

The ultimate goal of this research project was to predict which houses in the City had a likelihood of having a housing code violation. To make these predictions, the binary logistic regression equation was calculated for each house using the constant value and the variable coefficient values from the SPSS logistic analysis for the best performing model, which was the combined dataset of Areas 1, 2, and 3. Predicted probabilities were calculated using a logit transformation process because the logistic regression equation makes predictions using the log odds scale. First, the log (short for logarithm) odds value for each single-family home was calculated using the logistic regression equation where the constant was added to the products of the variable value and its corresponding coefficient (see Equation 1). Next, the log odds values

were multiplied by the exponential value (*e*), which is the inverse of a logarithm, so it mathematically converts the log odds (which is the logarithm of odds) to simple odds by cancelling the logarithm. Then, the odds values were converted to probabilities by dividing the odds value by 1 plus the odds value. This calculation determined the percentage of likelihood that a house had a violation (Simon 2013). Prediction percentages that were above 0.50 were given a value of "yes/1" and those below 0.50 were given a value of "no/0." The results were then mapped by joining the new dataset back to the parcels feature class for visual analysis.

The model validation process method was straight forward. Code enforcement continued to conduct their proactive survey in new areas of the City. One such area consisted of 376 homes in the Brentwood neighborhood near the geographic center of the City. The predicted values of the logistic regression model were cross-referenced with this area's actual survey data to determine how well the model actually performed. An overall accuracy percentage was determined by dividing the total number of correct predictions by the total number of actual proactive enforcement citations.

Chapter 4 Results and Predicted Violations

This chapter analyzes the outputs of both the non-spatial logistic regression models and the geographically weighted logistic regression (GWLR) models. Each iteration of the analysis over Areas 1, 2 and 3, as well as the five random samples and the combined areas produced differing statistics of varying strength and predictive capability. Section 4.1 looks specifically at the logistic regression model, how multicollinearity was minimized, the predictive capability of the models, and each models' statistical significance. Section 4.2 describes how each iteration of the model was affected by incorporating geographic variability using the GWR4 software and whether or not the logistic regression model was improved with GWLR. Section 4.3 discusses the reasons why the final prediction model was chosen, and Section 4.4 discusses the model validation using observed data from field inspections that were collected after the initial data collection of this research project. It also provides predictions for the 29,000 single-family homes in Victorville that were not part of the model training.

4.1 Binary Logistic Regression Results

Results of the logistic regression models were analyzed using UCLA's Institute for Digital Research and Education (IDRE) Annotated SPSS Output Logistic Regression resource. This is a free online resource center where annotated statistical interpretation instructions can be found for several different kinds of statistical outputs, including logistic regression. The following section discusses the key statistics in the SPSS output results, including the Wald chi-squared test used to determine if the constant is statistically significant from zero, the score and significance test used to determine if an independent variable was a good predictor, the test for overall model significance given by the chi-square statistics and its p-value, and the pseudo R squared values used to interpret model goodness-of-fit. The overall percentage value from the

Step1 Classification table is of importance because it gives the percentage of cases that were correctly classified by the logistic regression model. Furthermore, the Wald and significance test were analyzed to determine if the coefficient values for each variable were statistically significant from zero, allowing the null hypothesis to be rejected. The output also contains the coefficient values for each variable and the constant value that were used to create the logistic regression prediction equation. An example of the SPSS output and which statistics were of importance for the purposes of this research project can be found in Appendix B.

4.1.1. Interpretation of Results

This section discusses in detail how the results of the logistic regression output of SPSS in the context of the Area 1 values for reference. Each table shown in this section contains the portions of the logistic regression output that were of importance as previously explained. The tables are followed by a discussion on what these outputs mean, how they equate to the strength of the model and the relationships between the dependent and independent variables.

Interpretation began with the multicollinearity test of the variables. These tests showed that the code case, total value, and near distance variables had VIF values greater than 3 meaning that multicollinearity was present. Code case was removed because it provided the least amount of information compared to the case count, days to previous violation, and nearby cases variables that it was collinear with. Total value was removed because it was not normalized to the structure size where value per square foot was normalized. Near distance was removed because the network distance variable is a more accurate representation of reality. Residents cannot pass through walls or structures and would likely travel in a vehicle to a liquor store. Once these variables were removed, the linear regression test of the independent variables showed little multicollinearity with VIF values all below 3. Table 3 shows the final iteration of the

multicollinearity test for Area 1 while the remainder of the multicollinearity tests for the other iterations of the model can be found in Appendix C.

Table 3: Area 1 multicollinearity test

VARIABLE	VIF VALUE
Property Size (LOTSQFT)	1.519
Floor Area Ratio (FLOOR_AREA)	2.547
Number of Building Stories (NOSTORY_DUM)	1.908
Length of Ownership (LENGTH_OWN)	1.245
Structure Age (STRUCT_AGE)	1.019
Value Per Square Foot (VALUE_PSF)	1.306
Days to Previous Violation (DAYS_TO_VI)	2.593
Number of Cases 2005 to Present (CASE_COUNT)	2.664
Tax Default Status (TAX_STATUS_DUM)	1.040
Occupancy Type (OCCUPANCY_DUM)	1.185
Network Distance to Liquor Store (NETWORK_DI)	1.169
Number of Nearby Cases (NEARBY_CAS)	1.306

The variables that did not show multicollinearity were then input into the logistic regression model for Area 1 which yielded non-significant independent variables in the form of structure age, value per square foot, tax status, occupancy, and corporation owned. Removal of these variables did not affect the statistical significance of the predictive power of the other variables because each one is analyzed separately within SPSS. Once non-significant variables were removed, the remaining variables were statistically significant at the 0.05 level, upholding the assumption that the logistic regression model is fit using only significant variables. Table 4 shows the variable significance table before the non-significant variables were removed and Table 5 shows the variables that remained after the non-significant variables were removed.

Table 4: Area 1 variable selection with non-significant variables

VARIABLE	SIGNIFICANCE
Property Size (LOTSQFT)	.022
Floor Area Ratio (FLOOR_AREA)	.000
Number of Building Stories (NOSTORY_DUM)	.000
Length of Ownership (LENGTH_OWN)	.009
Structure Age (STRUCT_AGE)	.208
Value Per Square Foot (VALUE_PSF)	.698
Days to Previous Violation (DAYS_TO_VI)	.000
Number of Cases 2005 to Present (CASE_COUNT)	.000
Tax Default Status (TAX_STATUS_DUM)	.208
Occupancy Type (OCCUPANCY_DUM)	.934
Network Distance to Liquor Store (NETWORK_DI)	.000
Number of Nearby Cases (NEARBY_CAS)	.000
Corporation Owned (CORP_OWNED)	.504

Table 5: Area 1 remaining variables once non-significant variables were removed

VARIABLE	SIGNIFICANCE
Property Size (LOTSQFT)	.022
Floor Area Ratio (FLOOR_AREA)	.000
Number of Building Stories (NOSTORY_DUM)	.000
Length of Ownership (LENGTH_OWN)	.009
Days to Previous Violation (DAYS_TO_VI)	.000
Number of Cases 2005 to Present (CASE_COUNT)	.000
Network Distance to Liquor Store (NETWORK_DI)	.000
Number of Nearby Cases (NEARBY_CAS)	.000

The null model was then analyzed by looking at the p-value of the constant and the overall percentage value. As Table 6 shows, the p-value of the constant (or intercept) was .000 meaning that it was statistically significant and the null hypothesis could be rejected.

Table 6: The null model significance test

	B	Standard Error	Wald	Deg. Freedom	Sig.	Exp(B)
Constant	-1.119	.096	135.078	1	.000	.326

In SPSS, the null model (which excludes independent variables) predicts all values to be "no." The overall percentage value in the output is the number of predicted "no" values that were

correct in the dependent variable data. This essentially shows how many "no" values and how many "yes" values are in the training dataset, which in the case of Area 1 was 75.4% "no" values. The percentage was calculated by dividing the number of correct "no" values, 438, by the total cases in the dataset, 581. The remaining 143 cases had a value of "yes" because the null model predicted them incorrectly. Table 7 below, which is the output from SPSS, shows the total "no" values, the total "yes" values, and the overall accuracy percentage of the null model. Essentially, if independent variables were not included in the logistic regression equation, the model would be able to accurately predict 75.4% of the cases in the dataset. This percentage value became the benchmark against which the regression model was evaluated once the variables were input for this particular model.

Table 7: The null model predictions

	Predicted		
Observed	NO	YES	Percentage Correct
PROACTIVE NO	438	0	100.0%
PROACTIVE YES	143	0	0.0%
Overall Percentage			**75.4%**

The next output analyzed was the model significance test, which is indicated by the Omnibus Tests of Model Coefficients table as shown in Table 8. Here, the *model* entry at the bottom of the table indicates the statistical significance of the model. The chi-square value of 154.096 and p-value of .000 indicate that this is a significant model because the significance threshold is the .05 level. The degrees of freedom column was not used to interpret the logistic regression model using the *Enter* method, as is only of value in a stepped model in SPSS.

Table 8: Area 1 omnibus model significance test

	Chi-Square	Deg. Freedom	Sig.
Step	154.096	8	.000
Block	154.096	8	.000
Model	154.096	8	.000

The pseudo R squared values in the Model Summary Table were then analyzed. These R squared values cannot be interpreted in the same manner as ordinary least squares (OLS) because they are not calculated the same. The Cox and Snell R squared value is always calculated to be on a scale from 0 to 0.75, and the Nagelkerke R squared value calculates an adjustment to the Cox & Snell R squared to place it on a range from 0 to 1 in order to give it the appearance of an OLS R squared value, which is easier to interpret. IDRE advises careful interpretation of these values, but the intent is to show how much of the variation in the dependent variable is explained by the independent variables. According to Clark and Hosking (1986), a pseudo R squared value above 0.20 for Cox & Snell indicates a model is acceptable. These values are shown in Table 9.

Table 9: Area 1 pseudo R squared values

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	494.347	.233	.346

The Variables in the Equation output table shows the model coefficients (which are on the log odds scale), the Wald and significance test of the coefficients, and the odds ratio. These statistics are primarily used for interpreting how each independent variable affects the model in terms of positive and negative correlations. For example, if the odds ratio of the variable is less than 1, an increase in the value of that variable will cause a decrease in the odds of the event occurring, which is an inverse relationship. If the odds ratio is greater than 1, this indicates that an increase in the value of the variable will cause the odds of the dependent variable occurring to

increase, which is a direct relationship. If the odds ratio is exactly 1, there is no relationship between the dependent variable and the corresponding dependent variable. In SPSS, the software automatically rounds the odds ratio to three significant figures. The option to increase the number of significant figures is available. In this research project, the number of significant figures was increased to six. This caused odds ratios that were exactly 1.000 at three significant figures to become either slightly less than 1 or slightly more than 1, indicating that there is a very slight relationship between the dependent and independent variables in all iterations of the model.

For the floor area variable, the odds ratio is .019 meaning that if the floor area is increased, the odds of having a code enforcement violation decrease in Area 1. Since the Wald test is significant at the 0.05 level (the p-value is .025), the null hypothesis that the coefficient is equal to zero can be rejected. For the lot size, length of ownership, days to violation, and travel distance variables, the odds ratio was 1.000 when rounded to three decimal places. The output table was modified to show odds ratios to six decimal places, revealing a very small value for each of these variables. This indicated that there was a very slight inverse or direct relationship to the dependent variable, depending on the value of the odds ratio. The odds ratios and significance tests for the remaining variables are shown in Table 10.

It is worth noting here that although interpreting the odds ratios is interesting and provides much information about the model, it is not a vital factor in making predictions, which is the purpose of this research project. The primary focus for the Variables in the Equation output table was on the variable coefficient values and the value of the constant, both shown in the column labeled "B." These values are used in forming the logistic regression equation to make predictions. Interpretation of a .000 coefficient value is given in Chapter 5 as it relates to

the scale of the units of measurement of the variable.

Table 10: The coefficient and odds ratio output table with variables in the equation

Variable	B	Std. Error	Wald	d.f.	Sig.	Exp(B)
Lot Size (LOTSQFT)	.000	.000	.931	1	.335	1.000
Floor Area (FLOOR_AREA)	-3.978	1.772	5.040	1	.025	.019
Number of Stories (NOSTORY)	-.140	.326	.185	1	.667	.869
Length of Ownership (LENGTH_OWN)	.000	.000	4.281	1	.039	1.000
Days to Previous Violation (DAYS_TO_VI)	.000	.000	22.152	1	.000	1.000
Number of Previous Cases (CASE_COUNT)	.147	.115	1.645	1	.200	1.158
Travel Distance to Liquor Store (NETWORK_DIST)	.000	.000	20.211	1	.000	1.000
Number of Nearby Cases (NEARBY_CASE)	.059	.033	3.304	1	.069	1.061
Constant	2.396	1.242	3.718	1	.054	10.976

The final output table is the Block 1 Classification Table shown in Table 11. This table provides the most important piece of information in the output. This table shows how well the model performed with the variables included. Again, for Area 1, the null model with no variables included was able to correctly predict 75.4% of the cases. With the variables included, the model was able to correctly predict 84.3% of the cases. This is an increase of nearly 10%, meaning that this is a good predictive model and performed well against the null model. The *cut value* of .500 simply indicates that if the predicted probability of a case was below that cut value, it was given a value of "no" and if the predictive probability was above that cut value, it was given a value of "yes." If more certainty was to be given to the predicted probabilities, this value could be increased to .600, which would in turn cause borderline predictions around a value of .500 to be given a value of "no" in the model and only the stronger predicted probabilities (those

above .600) would be given a value of "yes."

Table 11: The overall number of cases predicted correctly for Area 1

	Predicted		
Observed	NO	YES	Percentage Correct
PROACTIVE NO	410	28	93.6%
PROACTIVE YES	63	80	55.9%
Overall Percentage			**84.3%**

4.1.2. Logistic Regression Model Results

This section contains the results of the logistic regression models for Areas 1, 2, and 3, the five random samples, and the combined areas. The summary tables contained in this section show the most interesting results that pertain to the interpretation of the logistic regression models. Each table is followed by a discussion on the key differences between the various iterations of the model. Chapter 5 discusses what these results mean and what findings can be made from them about the data. To begin, Table 12 shows the key statistics of the iterations of the logistic regression model used to compare and make conclusions about the strength of each model.

Table 12: The results of the model iterations and key statistics

Model	Null Model Constant	Sig.	Null Model Accuracy	Chi-square Value	Sig.	Cox & Snell R Squared	Nagelkerke R Squared	Prediction Model Accuracy	Accuracy Difference	Model Assessment
Area 1	-1.119	.000	75.4	154.096	.000	0.233	0.346	84.3	8.9	Good
Area 2	-0.371	.001	59.2	124.989	.000	0.327	0.441	76.9	17.7	Very Good
Area 3	-0.261	.000	56.5	274.345	.000	0.190	0.255	71.1	14.6	Poor
Sample 1	-0.464	.000	61.4	99.047	.000	0.180	0.244	72.6	11.2	Poor
Sample 2	-0.456	.000	61.2	126.419	.000	0.223	0.303	76.2	15.0	Good
Sample 3	-0.456	.000	61.2	111.749	.000	0.200	0.272	73.2	12.0	Poor
Sample 4	-0.405	.000	60.0	119.855	.000	0.213	0.288	75.2	15.2	Good
Sample 5	-0.498	.000	62.2	133.114	.000	0.234	0.318	77.4	15.2	Good
Combined	-0.484	.000	61.9	506.853	.000	0.206	0.280	74.9	13.0	Good

In all of the models, the null model constant value was negative and they each had a significance value of 0.000, so the null hypothesis that the model's constant was equal to zero could be rejected at the 0.01 level. Area 1 had the highest null model accuracy percentage and was approximately 15% greater than the other models. In every model except the combined areas, the chi-squared value was fairly low and every model's chi-squared significance value was 0.000 indicating that these were all statistically significant models. The pseudo R squared values varied with Random Sample 1 having the lowest value and Area 2 having the highest value.

The model assessment column in Table 12 indicates if the resulting model for each area was good, very good, or poor. This assessment was based on the combination of the model's pseudo R squared values, the significance value of the omnibus test of model coefficients which is indicated by the chi-squared value and corresponding significance value, and the change in the predicted accuracy from the null model to the prediction model. To be classified as a good model, the omnibus test had to yield a low chi-squared value with a significance value less than 0.05, the pseudo R squared values had to be greater than 0.20, and there had to be an increase in the prediction accuracy from the null model to the prediction model.

The accuracy of the prediction model was the most important piece of information to take from this table. Area 1 produced the highest overall value but had the lowest accuracy difference over its corresponding null model. However, this was a good model because the pseudo R squared values were above 0.20 and the overall accuracy percentage was high. The best model came in Area 2 because it had the highest pseudo R squared values and the greatest change in accuracy over the null model. Area 3 along with Random Samples 2 and 3 were the weakest models because the pseudo R squared values were at or below the 0.20 threshold and the change

in accuracy percentages were low. The remaining models all had pseudo R squared values above 0.20 and the change in accuracy was 15% or higher, meaning that these were acceptable models.

4.1.3. Logistic Regression Coefficients

This section contains a brief discussion of dependent and independent variable relationships and what it means for the corresponding regression equation coefficients. Table 13 contains the results of the coefficient calculations in SPSS, which are on the log odds scale, the converted odds ratio used for interpretation, and the significance value of the odds ratio. The null hypothesis here is that the coefficient value is equal to zero and the statistical significance level to reject this null hypothesis was the 0.05 level. As previously explained, the odds ratio determines the likelihood of the dependent variable occurring with a change in the independent variable in either the positive or negative direction.

Table 13: Results of the coefficient calculations, the odds ratio, and significance for the logistic regression models

Variable Name and (Alias)	Coefficients & Relationships		
	Variable Coefficient	Variable Odds Ratio - Exp(B)	Odds Ratio Significance
AREA 1			
Lot Size (LOTSQFT)	0.000	1.000	0.335
Floor Area (FLOOR_AREA)	-3.978	0.019	0.025
Number of Stories (NOSTORY)	-0.140	0.869	0.667
Length of Ownership (LENGTH_OWN)	0.000	0.039	1.000
Value per Sq Ft (VALUE_PSF)	0.001	1.001	0.928
Days to Previous Violation (DAYS_TO_VI)	0.000	1.000	0.000
Number of Previous Cases (CASE_COUNT)	0.147	0.200	1.158
Travel Distance to Liquor Store (NETWORK_DIST)	0.000	1.000	0.000
Number of Nearby Cases (NEARBY_CASE)	0.059	1.061	0.069
Constant	2.396	10.976	0.054

Variable Name and (Alias)	Coefficients & Relationships		
	Variable Coefficient	Variable Odds Ratio - Exp(B)	Odds Ratio Significance
AREA 2			
Value per Sq Ft (VALUE_PSF)	0.001	1.001	0.928
Days to Previous Violation (DAYS_TO_VI)	-0.001	0.999	0.000
Number of Previous Cases (CASE_COUNT)	0.067	1.069	0.695
Number of Nearby Cases (NEARBY_CASE)	0.001	1.001	0.973
Constant	1.261	3.527	0.154
AREA 3			
Length of Ownership (LENGTH_OWN)	0.000	1.000	0.109
Days to Previous Violation (DAYS_TO_VI)	-0.001	0.999	0.000
Number of Previous Cases (CASE_COUNT)	-0.250	0.779	0.033
Occupancy (OCCUPANCY)	0.071	1.074	0.668
Cartesian Distance to Liquor Store (NEAR_DIST)	0.000	1.000	0.201
Travel Distance to Liquor Store (NETWORK_DIST)	0.000	1.000	0.015
Number of Nearby Cases (NEARBY_CASE)	0.034	1.034	0.140
Constant	2.114	8.284	0.000
RANDOM SAMPLE 1			
Days to Previous Violation (DAYS_TO_VI)	0.000	1.000	0.000
Number of Previous Cases (CASE_COUNT)	-0.016	0.984	0.919
Number of Nearby Cases (NEARBY_CASE)	0.060	1.062	0.035
Constant	0.351	1.421	0.369
RANDOM SAMPLE 2			
Number of Stories (NOSTORY)	0.839	2.315	0.000
Days to Previous Violation (DAYS_TO_VI)	-0.001	0.999	0.000
Number of Previous Cases (CASE_COUNT)	-0.168	0.846	0.210
Cartesian Distance to Liquor Store (NEAR_DIST)	0.000	1.000	0.926

Variable Name and (Alias)	Coefficients & Relationships		
	Variable Coefficient	Variable Odds Ratio - Exp(B)	Odds Ratio Significance
Travel Distance to Liquor Store (NETWORK_DIST)	0.000	1.000	0.058
Number of Nearby Cases (NEARBY_CASE)	0.016	1.016	0.637
Constant	1.225	3.405	0.034
RANDOM SAMPLE 3			
Floor Area (FLOOR_AREA)	-2.158	0.116	0.094
Number of Stories (NOSTORY)	0.399	1.491	0.112
Structure Age (STRUCT_AGE)	0.007	1.007	0.359
Days to Previous Violation (DAYS_TO_VI)	0.000	1.000	0.000
Number of Previous Cases (CASE_COUNT)	-0.091	0.913	0.446
Cartesian Distance to Liquor Store (NEAR_DIST)	0.000	1.000	0.685
Travel Distance to Liquor Store (NETWORK_DIST)	0.000	1.000	0.009
Number of Nearby Cases (NEARBY_CASE)	0.019	1.019	0.554
Constant	2.119	8.324	0.011
RANDOM SAMPLE 4			
Number of Stories (NOSTORY)	0.598	1.818	0.008
Value per Sq Ft (VALUE_PSF)	-0.017	0.983	0.005
Days to Previous Violation (DAYS_TO_VI)	0.000	1.000	0.000
Number of Previous Cases (CASE_COUNT)	0.012	1.012	0.938
Occupancy (OCCUPANCY)	-0.168	0.845	0.510
Travel Distance to Liquor Store (NETWORK_DIST)	0.000	1.000	0.013
Number of Nearby Cases (NEARBY_CASE)	0.061	1.063	0.039
Constant	2.111	8.254	0.004
RANDOM SAMPLE 5			
Floor Area (FLOOR_AREA)	-0.479	0.620	0.727
Number of Stories (NOSTORY)	0.602	1.826	0.025
Days to Previous Violation (DAYS_TO_VI)	-0.001	0.999	0.000
Number of Previous Cases (CASE_COUNT)	-0.158	0.854	0.320
Occupancy (OCCUPANCY)	-0.170	0.844	0.518

Variable Name and (Alias)	Coefficients & Relationships		
	Variable Coefficient	Variable Odds Ratio - Exp(B)	Odds Ratio Significance
Cartesian Distance to Liquor Store (NEAR_DIST)	0.000	1.000	0.864
Travel Distance to Liquor Store (NETWORK_DIST)	0.000	1.000	0.101
Number of Nearby Cases (NEARBY_CASE)	0.055	1.056	0.086
Constant	1.414	4.111	0.110
COMBINED AREAS			
Floor Area (FLOOR_AREA)	-0.623	0.536	0.309
Number of Stories (NOSTORY)	0.437	1.548	0.000
Days to Previous Violation (DAYS_TO_VI)	0.000	1.000	0.000
Number of Previous Cases (CASE_COUNT)	-0.088	0.916	0.203
Cartesian Distance to Liquor Store (NEAR_DIST)	0.000	1.000	0.659
Travel Distance to Liquor Store (NETWORK_DIST)	0.000	1.000	0.000
Number of Nearby Cases (NEARBY_CASE)	0.022	1.022	0.143
Constant	1.602	4.961	0.000

The most important thing to note here is a variable with a coefficient of 0.000 does not impact the outcome of the logistic regression equation. It causes the associated variable to drop out of the equation because the product of the coefficient value and the variable value are zero. Since the logistic regression equation is the addition of the constant value with the products of the coefficient and variable values, adding a zero value does not change the outcome because any number when added to zero is still the same. That being said, many of the variables in these regression models could not reject the null hypothesis of the coefficient equaling zero. This detracts from the overall strength of the model, which is discussed later in Chapter 5.

In all iterations of the model, the days to previous violation variable was a significant predictor and had a statistically significant odds ratio at or very close to 1.000. This means that

there was an issue with this variable. The number of significant digits of the output tables was increased from 3 to 6, which revealed that the coefficients of this variable did in fact have a value, albeit extremely small. So here, the null hypothesis that the variable's coefficient is zero can be rejected even though the odds ratio shows no positive or negative relationship. In addition, the null hypothesis for the constant in each model could only be rejected for the combined areas, Area 3, and Random Samples 2, 3 and 4. Also note that in four of the six instances where the number of building stories was a significant predictor, it also showed a significant relationship between the independent and dependent variables.

One will also notice the absence of the corporation owned yes/no and tax status variables in all iterations of the model, meaning that these variables contributed nothing to the logistic regression model training and showed no relationships between the dependent and independent variables. Occupancy only appeared twice and structure age appeared only once in the model iterations, which is discussed further in Chapter 5.

4.2 GWR 4 Results

The following section discusses the results of performing the geographically weighted logistic regression analysis using the GWR4 software. Areas 1, 2, and 3, the five random samples, and the combined areas were input into the GWR4 software using only the remaining variables from the logistic regression models from SPSS to determine if there was any effect of location on the data. For all but one of the model iterations, the local geographical weighting model improved upon the global model for goodness of fit to some extent. Furthermore, the geographical weighting was able to increase the model prediction percentage, although, many of these increases were small at approximately one percentage point.

The next section describes how the results of the GWR4 software were interpreted in the context of Area 1's data for a reference. Interpretations were based off of the GWR4 user manual, which explains how to perform the model, what the different tool selections mean, suggestions on which kernel types to use based on the model type, and what each of the output statistics means. This manual is attached to the GWR4 download file from Arizona State University.

4.2.1. Interpretation of GWR4 Results

The GWR4 software first conducts a "global" regression analysis where local variation in the data is not accounted for. This is similar to, but not the same as the logistic regression calculations of SPSS; the SPSS outputs give more information and were used to removed non-significant predictors from the model. The global regression of GWR4 is not important for making predictions, but is still worth noting because the *percent of deviance explained* value was used to compare the global and local regression models of GWR4. Also, GWR4 produces an Akaike Information Criterion (AIC) value that is used to compare the fit with different models. Essentially, the smaller the AIC, the better the model fit. For Area 1, the percent deviance explained value was 0.23 and the AIC was 503.116. Percent of deviance explained can be interpreted in the same manner as the pseudo R squared values of SPSS. The next output was the optimal bandwidth distance as computed by the *golden bandwidth selection search* method of GWR4. Chapter 3 explained why the golden bandwidth selection method was chosen to determine the optimal bandwidth measure, but it essentially determines where the spatial variability is most pronounced in the data. This bandwidth is given as the number of nearest neighbors and not in a distance. This is because the kernel type is adaptive, meaning the actual distance between the analyzed cases to the furthest point in the kernel will vary. The bandwidth

value is the number of cases that fall within the adaptive kernel that produces the optimal amount of influence on the case being analyzed, according to GWR4. For Area 1, this was the 488 nearest neighbors out of a total of 581 different cases.

The next output of GWR4 was the local regression model where spatial relationships are considered. The percent of deviance explained in Area 1 when spatial variation was accounted for was 0.27, which is an increase over the global model. It also produced coefficient values based on several different aggregation methods, which are not important for this research project. The *y* and *yhat* comparison for Area 1 produced an overall correct prediction percentage of 84.9%, which is a very slight increase from the logistic regression percentage of 84.3%. Here, the local model out-performed the global model by increasing the percent of deviance explained value, but the overall accuracy did not increase significantly. Figure 5 shows the results of the GWR4 global model calculated by the software, Figure 6 shows the optimum kernel bandwidth, and Figure 7 shows the results of the local model. A full GWR4 output and an example of the listwise table showing the *y* and *yhat* values can be found in Appendix D.

```
*************************************************************************
  Global regression result
*************************************************************************
  < Diagnostic information >
Number of parameters:                  9
Deviance:                        494.346944
Classic AIC:                     512.346944
AICc:                            512.662180
BIC/MDL:                         551.629701
Percent deviance explained         0.237640

Variable                  Estimate    Standard Error      z(Est/SE)        Exp(Est)
------------------- --------------- --------------- --------------- ---------------
Intercept                 2.255239        1.074126        2.099604        9.537570
LOTSQFT                   0.000055        0.000057        0.964632        1.000055
FLOOR_AREA               -3.977777        1.771890       -2.244935        0.018727
NOSTORY_DUM               0.140443        0.326199        0.430543        1.150783
LENGTH_OWN                0.000147        0.000071        2.069021        1.000147
DAYS_TO_VI               -0.000373        0.000079       -4.706586        0.999627
CASE_COUNT                0.146921        0.114551        1.282580        1.158263
NETWORK_DI               -0.000321        0.000072       -4.495662        0.999679
NEARBY_CAS                0.059460        0.032712        1.817709        1.061264
```

Figure 5: Global regression result for Area 1 in GWR4

```
  Bandwidth and geographic ranges
Bandwidth size:                  488.233329
Coordinate                Min               Max             Range
--------------- --------------- --------------- ---------------
X-coord            -117.329850     -117.318008        1.084536
Y-coord              34.543607       34.556333        1.415133
 (Note: Ranges are shown in km.)
```

Figure 6: Area 1 bandwidth selection

```
  Diagnostic information
Effective number of parameters (model: trace(S)):                          16.671441
Effective number of parameters (variance: trace(S'WSW^-1)):                12.496829
Degree of freedom (model: n - trace(S)):                                  564.328559
Degree of freedom (residual: n - 2trace(S) + trace(S'WSW^-1)):            560.153946
Deviance:                              468.727419
Classic AIC:                           502.070302
AICc:                                  503.116258
BIC/MDL:                               574.836988
Percent deviance explained               0.277149

*********************************************************
 << Geographically varying (Local) coefficients >>
*********************************************************
Estimates of varying coefficients have been saved in the following file.
    Listwise output file: C:\Users\Matthew\Documents\THESIS PROJECT 2\FULL_DATA_OUTPU

Summary statistics for varying (Local) coefficients
Variable                        Mean             STD
-------------------- --------------- ---------------
Intercept                   2.693315        0.746112
LOTSQFT                     0.000056        0.000057
FLOOR_AREA                 -2.557080        2.853220
NOSTORY_DUM                 0.052745        0.211965
LENGTH_OWN                  0.000156        0.000032
DAYS_TO_VI                 -0.000360        0.000122
CASE_COUNT                  0.124512        0.050806
NETWORK_DI                 -0.000392        0.000118
NEARBY_CAS                  0.019748        0.026708
```

Figure 7: Local model output for Area 1

4.2.2. GWR4 Model Results

This section contains the results of the GWR4 model iterations. Table 14 summarizes these results beginning with the global regression model's AIC value and deviance explained value. That is then followed by the bandwidth distances selected by the *golden bandwidth selection search* of GWR4 and the corresponding total number of cases in each model. It also shows the outputs of the local regression model, including the AIC value and percentage of deviance explained when spatial relationships are considered. The remaining portion of the table shows the accuracy comparisons between the logistic regression models of SPSS and those of the GWR4 software.

Table 14: GWR4 model iteration summary table containing important statistics

MODEL	Global Model		Bandwidth		Local Model		Accuracy		
	AIC	% Deviance Explained	Total Cases in Model	Optimum Bandwidth	AIC	% Deviance Explained	Logistic Model	GWLR Model	Diff.
Area 1	512.3469	0.23	581	488	502.0703	0.28	84.3%	84.9%	0.6%
Area 2	312.3737	0.29	316	216	313.3339	0.33	76.9%	80.1%	3.2%
Area 3	1523.20885	0.15	1301	954	1513.80266	0.17	71.1%	72.2%	1.1%
RS 1	575.878	0.14	500	500	559.1855	0.19	72.6%	73.2%	0.6%
RS 2	555.4258	0.18	500	500	540.761	0.23	76.2%	75.6%	-0.6%
RS 3	574.0514	0.17	500	500	564.7624	0.21	73.2%	74.6%	1.4%
RS 4	569.1571	0.18	500	500	550.7589	0.24	76.0%	75.6%	-0.4%
RS 5	547.9625	0.20	500	500	535.4414	0.26	77.4%	77.8%	0.4%
Combined	2431.0602	0.17	2198	897	2336.4292	0.23	74.9%	76.3%	1.4%

Each iteration of the GWR4 model decreased the AIC value of the global model in the local model with the exception of Area 2 where it was increased by 1. This indicates that the GWLR model fit better than the non-spatial logistic regression model everywhere except for Area 2. Note too that the percent of deviance explained was much lower for the GWR4 software than in the SPSS pseudo R squared values. This emphasizes the importance of using the SPSS software for performing the non-spatial regression calculations because it proved to be more accurate in terms of the percentage of variation explained in the model.

The optimum bandwidth selections were rather large compared to the total number of cases in each model. This is a strong indicator of the absence of spatial relationships in the data because these bandwidths encompass most, if not all of the cases in each study area; for each random sample, the bandwidth was equal to the number of cases. This means that the optimum bandwidth calculation, which is designed to pick up the point in the data where the spatial relationships are most pronounced, failed to find any sort of peak relationship point. Furthermore, the data sampling method employed for this research project may have been the cause of the large bandwidth sizes due to a lack of organization in the collection of the observed violation data. Unfortunately, this could not be controlled because the data was collected by City

code enforcement officers before the time of this research project. This is discussed further in Chapter 5 in the future work discussion.

The percent of deviance explained by the local model was increased over the global model in every iteration. However, this did not equate to a higher prediction accuracy for every iteration. For Random Samples 2 and 4, the accuracy of the predictions actually decreased. The rest of the accuracy percentages were approximately positive 1%, with the highest being in Area 2 at 3.2%. These increases are not very pronounced, and the implications of this are explained in Chapter 5 as well.

Spatial variation was also observed between the three study areas when analyzing the coefficient values of each variable in the GWLR output of the Combined Areas. For each study area, there was very low spatial variability among the data within each area, and each area differed from the other two. For example, the floor area variable had all data in Area 1 less than -0.50 standard deviations from the mean. In Area 2, all of the floor area data was in the -0.50 to 0.50 standard deviation range. Area 3 was the only study area to show variability in its data largely because of its larger sample size. Here, data points in the northwest section were less than -0.50 standard deviations, data points in the middle of the study area were between 0.50 and 1.7 standard deviations, and data points near the eastern edges of the study area were near the mean in the -0.50 to 0.50 standard deviation range. The absence of spatial variability in the data within each study area was likely caused by the large bandwidth size, which is discussed further in Chapter 5 along with the implications of the lack of variability in the GWLR output. Figure 8 below depicts the floor area coefficient analysis of the GWLR output for the Combined Areas. The remaining variable coefficient maps from the Combined Areas can be found in Appendix E.

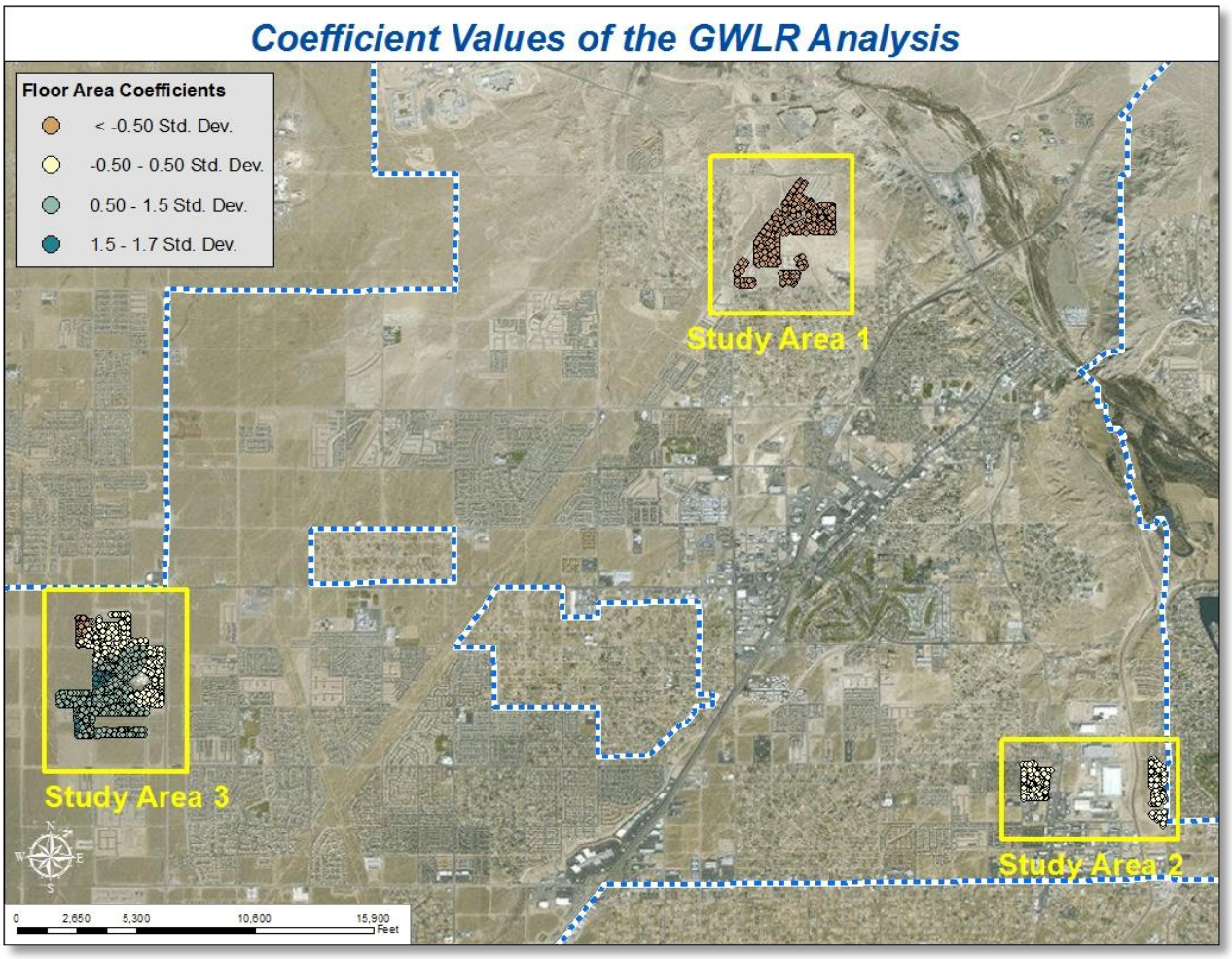


Figure 8: Floor area variable coefficient values of the GWLR analysis showing low variability.

4.3 Selection of the Prediction Model

The model iterations from logistic regression and geographically weighted logistic regression ranged from very good to poor when considering pseudo R squared values, percent of deviance explained, and overall prediction accuracy. This discussion explains the reasoning for why the final model used to make predictions for the entire City of Victorville was selected.

Of the non-spatial logistic regression models, Area 1 performed the best in terms of number of correctly predicted violations while Area 2 performed the best in terms of the amount of variation explained as shown in the pseudo R squared values of the model. These two models

would have been the top candidates had it not been for the neighborhood variation issue, which is explained in Chapter 5. Area 3 produced too weak of a model for consideration. All of the random samples were disregarded because the fluctuations in the significant predictor variables was too great, and the fluctuations in the strength of each sample's model would make predictions highly volatile.

The geographically weighted regression models were not considered for similar reasons. Area 1 and especially Area 2 were able to increase the accuracy of predictions when spatial relationships were considered, however, each area on its own did not encompass enough neighborhood variability to make good predictions for the entire city. Area 3 was also able to increase the accuracy, but its high AIC value and low percentage of deviance explained values were too low to be considered. The random samples were too inconsistent to be considered, especially since two of the samples actually reduced the accuracy of the model predictions. The combined areas in the spatial model produced an extremely high AIC value meaning that the model was not a good fit when spatial relationships were considered.

Therefore, the combined areas dataset of the non-spatial logistic regression models was the top candidate to make predictions across the entire city because it encompassed the most neighborhood variation within the data. It also had acceptable pseudo R squared values and it had three variables that had statistically significant variable relationships. Furthermore, the increase in the prediction accuracy over its null model was acceptable, and it contained the largest sample size. As a reminder, a large sample size is a primary assumption of the logistic regression technique.

4.4 Predictions

This final section of Chapter 4 looks at the predictions made by the combined areas non-spatial logistic regression model that was selected as explained in the previous section. The first part of this section discusses the predicted outcomes of the logistic regression equation for the single-family homes that were not part of the model training. The second section discusses how the model was validated by calculating the accuracy of the predictions with observed violations from field inspections that occurred after the data collection date of this research project, which was July 11, 2015.

4.4.1. Prediction of Violations

Following the logistic regression equation calculation and the log odds transformation process outlined in Section 3.3.5, predictions were made for all of the single-family homes in the City. Figure 9 shows a very small section of the City at the intersection of Bear Valley Rd and Amethyst Rd where there is a high density of homes. The red dots on the map indicate a housing code violation is more than 50% likely while a blue dot indicates that a violation is less than 50% likely. According to the model, 7,483 homes are likely to have a code violation. This is approximately 26% of the 29,000 single-family homes in the City. As a reference, if the number of predicted violations was divided evenly among the seven code enforcement officers in Victorville, they would each have to take on over 1,000 individual cases, which is a substantial workload.

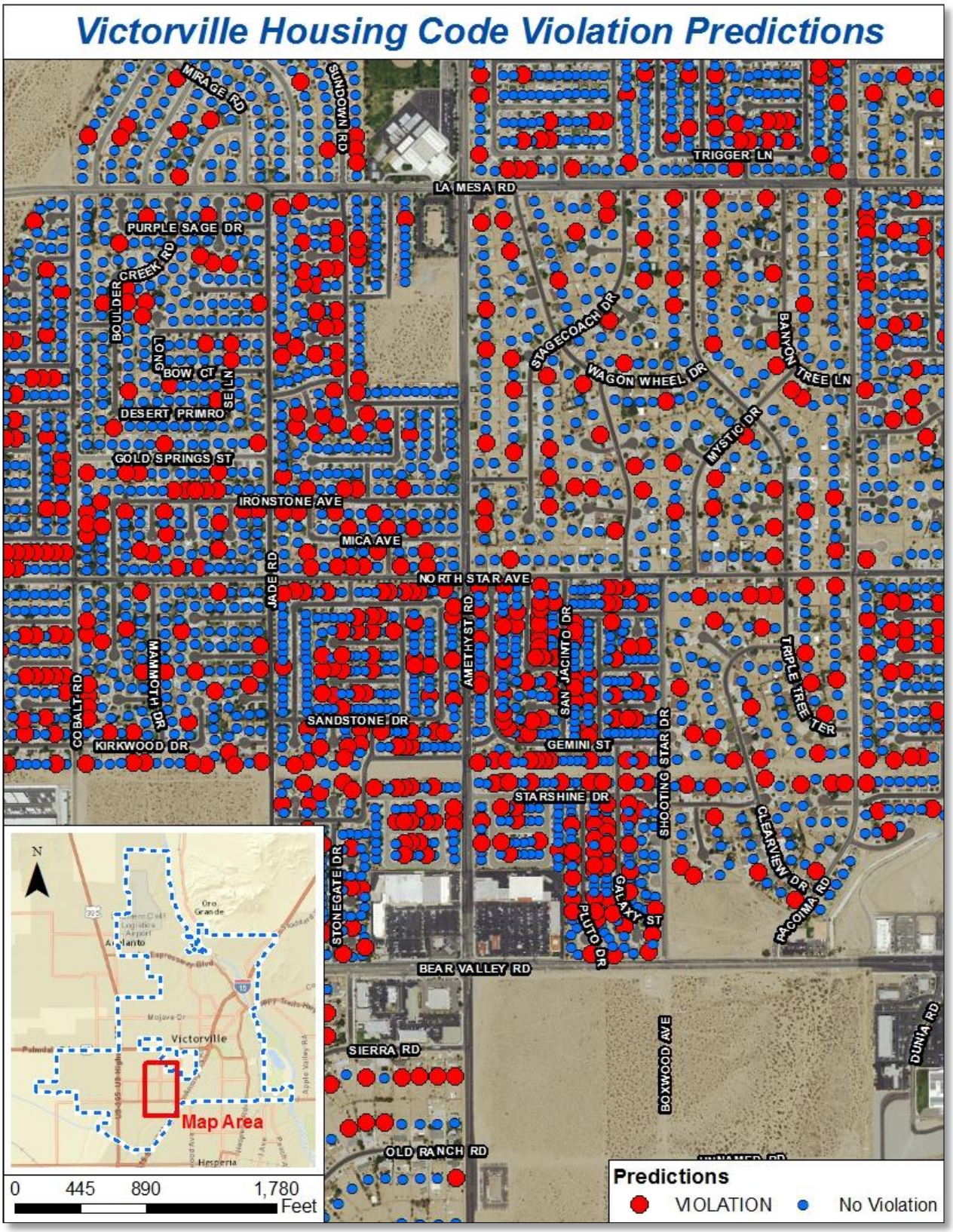


Figure 9: Predicted housing code violations based on logistic regression

Figure 10 shows an area of the City where there is a clear contrast between the predictions in four neighborhoods. Neighborhoods 1 and 2 outlined in red show a high density of probable housing code violations while Neighborhoods 3 and 4 show a very low density of probable housing code violations. This illustrates how neighborhoods can possibly be selected for proactive enforcement due to high likelihood of violations versus neighborhoods that can be overlooked in proactive code enforcement due to a very low likelihood of housing code violations.

Figure 10 also shows that the predictions can potentially depict differences among neighborhoods. Neighborhood 4 outlined in blue is a new housing tract that was built in the last 10 years. Neighborhood 3 is a tract of single story homes on very small lots, with a requirement that only seniors can be the primary resident. Neighborhoods 1 and 2 were both built in the mid-1990s and are likely beginning to show signs of decay. This appears to come across in the predictions, and is explained further in Chapter 5.

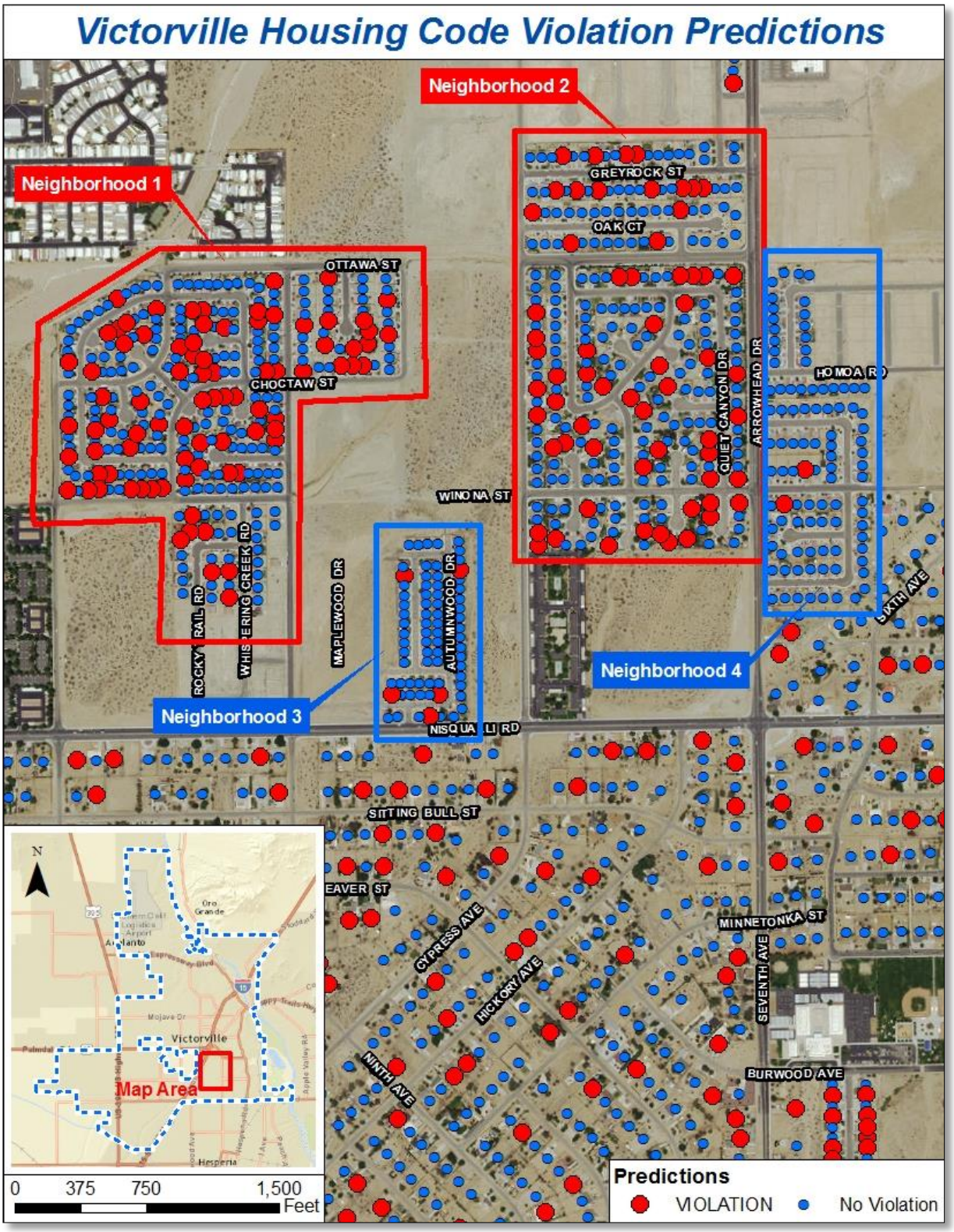


Figure 10: Neighborhood comparison of predictions

4.4.2. Model Validation

The neighborhood selected for model validation is shown in Figure 11. Again, it consisted of 376 single-family homes in the Brentwood neighborhood in the center of the City. The overall observed accuracy of the model was calculated in exactly the same manner as it was in the training model by comparing the number of correctly predicted "no" violations with the number of observed "no" violations, as well as the number of predicted "yes" violations with the number of observed "yes" violations. These were then totaled and compared against the total number of homes in the neighborhood. Table 15 shows the outcome of this calculation. The combined area training model calculated a 74.9% overall accuracy in SPSS. The observed accuracy from comparing the predicted with the observed indicated a 50.3% overall accuracy. This was calculated by adding the total number of correct "no" predictions and the total correct "yes" predictions and dividing that by the total number of cases, so (125 + 64) / 376. The implications of this outcome are discussed in detail in Chapter 5.

Table 15: Observed predicted accuracy of the model

Observed		**Predicted**		
	TOTAL	NO	YES	% Correct
NO	231	125	106	54.1%
YES	145	81	64	44.1%
TOTAL	376			**50.3%**

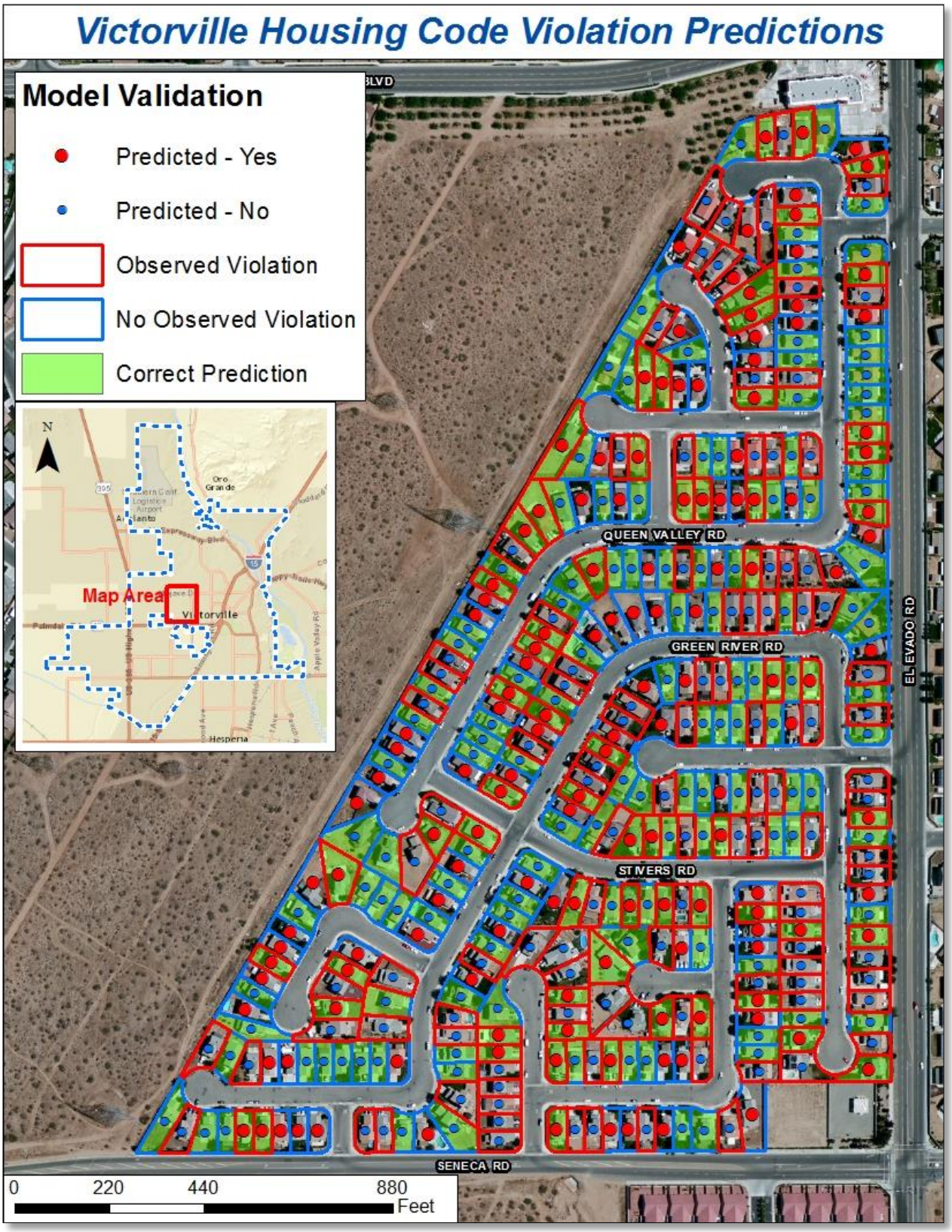


Figure 11: Model validation neighborhood

Chapter 5 Discussion and Conclusions

The results of the analysis proved to be quite interesting and many of the outcomes were unexpected. This chapter outlines the major findings from the non-spatial and spatial regression models and answers the research questions outlined in Chapter 1. In addition, this chapter discusses how this research project contributes to the work previously completed using a combined analysis methodology of logistic regression and geographically weighted regression as well as using regression modeling as a tool for making predictions for spatially occurring incidents, such as crime. Furthermore, this chapter discusses what the major limitations were in this research project and what work can be done in the future to improve upon its findings.

5.1 Findings

This section discusses the major findings of this research project. It illustrates what was of significance in the variable selection for each model, what relationships were found between the dependent and independent variables, what relationships could be found spatially using the GWLR technique, and the implications of the model predictions. It also discusses how each iteration of the model compared to the others and why the observed percentage of correct variables did not match up with the training model.

5.1.1. Non-spatial Findings

The multicollinearity tests revealed that the code case yes/no, lot sq ft, and total value variables showed collinearity in nearly all of the models. This outcome was actually expected because the code case yes/no variable simply showed if a code violation had been present in the past which is expressed in the days to previous violation and number of previous cases variables. Note here that the days to previous violation and case count variables did not show collinearity

because they were on different scales of measurement; days to previous violation was a length of time and case count was a tally, meaning they tell different stories. The total value variable was normalized by dividing it by the square footage of the home, which was the likely cause of the collinearity between it and the value per sq ft variable. These were both dollar measurements of the value, and once total value was removed for the reasons expressed in Chapter 4, the VIF value of the value per sq ft variable dropped to approximately 1.3 which means there was little to no collinearity with other variables. Finally, the lot square footage variable expressed collinearity with the floor area variable because the floor area variable was the building square footage divided by the lot square footage, meaning that the lot square footage information was also expressed in the floor area variable. Again, once the lot square footage variable was removed, the VIF of floor area dropped to below 2.0 indicating little multicollinearity. The exception to this was in Areas 1 and 2 where the near distance and network distance variables showed collinearity with each other. This was likely because Area 1 and 2 were the smaller sample sizes in terms of geographic area covered, so the near and network distances were similar. Area 3, the combined areas, and the random samples encompassed larger geographic areas giving these distances more variability. Near distance was removed as explained in Chapter 4 and the VIF value of network distance dropped below 2.0 in both Areas 1 and 2.

The nine iterations of the non-spatial logistic regression model produced results that were unique to each sample of data. As Chapter 4 illustrated, each training model had a different set of significant predictor variables. For example, Area 1 retained lot size, floor area, number of stories, length of ownership, value per sq ft, days to previous violation, previous case count, network distance, and number of nearby cases while Area 2 only retained value per sq ft, days to previous violation, previous case count, and number of nearby cases as significant predictors.

This was likely caused by variability in the data between each of the study areas. As explained in Chapter 3, each study area where proactive enforcement inspections were conducted had different neighborhood characteristics. Area 1 had a homeowner's association, Area 2 had a very low income neighborhood and a high income neighborhood, and Area 3 was somewhere in the middle of these. The neighborhood characteristics, which were not accounted for in this research project, likely caused the differences in which predictors were significant for each area.

Furthermore, variables that were initially thought to be good predictors of code violations were not significant predictors at all. For example, the corporation owned variable was not a significant predictor in any of the models. The initial thought was that if a property was owned by a corporation, it was likely to be rented which would create occupancy turn over and lead to a higher likelihood of a violation. According to the logistic regression models, this was resoundingly not the case. In addition, structure age only appeared as a significant predictor in one of the models. The initial thought with this variable is that the structure age and the likelihood of a violation would be a direct relationship, meaning that as age increases, the chance of a violation also increases because the older homes must be maintained and if the occupant does not make these repairs, code violations appear. These models indicate that the structure age is not a significant determinant of code violations. Another variable that was thought to be a significant predictor was the tax default variable, which did not appear in any of the regression models. Here, the reasoning was that if an owner is defaulting on their property taxes, they are not maintaining their property because they cannot afford it. This was somewhat touched upon in Chapter 2 in the discussion that if there is a financial burden on an owner, caused by too much code enforcement or otherwise, the property is likely to not be maintained. This relationship may be more complicated than a simple binary variable and is worth exploring further.

The independent and dependent variable relationships also varied between the different iterations of the model. The most interesting item came with the days to violation variable where in most of the models, the coefficient value was near zero, the odds ratio was near 1.000, but the odds ratio was significant at the 0.01 level. It was discovered that the coefficient did not equal zero once the number of significant digits was increased from 3 to 6 in the output tables. These coefficient values were extremely small (0.000487 for example) but they still played a role in determining the outcome of the dependent variable. This was also the case with the network distance and length of ownership variables when they were significant predictors in the models. It was also found that the number of building stories had a positive relationship in the models where it was a significant predictor except for in Area 1 where it was a negative relationship, and in Random Sample 3 where the significance value was greater than 0.05 and the null hypothesis could not be rejected. This means that when the number of stories value increases from 0 (single story) to 1 (two story), the likelihood of having a code violation increases. Again, it was not a significant predictor in all the models, but this was still an interesting finding because the number of stories was not initially thought to have a significant relationship with code violations.

The variation in the strength of the different models was also interesting. In Area 2 where the sample size was smallest, the pseudo R squared values and model accuracy increase were the highest. Area 2 also had the second fewest number of significant predictors in the model at four. Random Sample 1 was the weakest model in terms of pseudo R squared values, but it performed better than Area 1 in terms of increasing the model's prediction accuracy. The five random samples showed some consistency in their null model accuracy with all percentages at approximately 61%. However, this was not reflected in the prediction model accuracy as two random samples performed lower than the other three. Also note here that the pseudo R squared

values vary from 0.18 to 0.234, meaning that there is inconsistency in model strength. This is an indication that there is possible variability in the neighborhoods where the data was collected; otherwise the random samples would have been more consistent in terms of model accuracy and strength because each sample would have had data that was similar.

In addition, neighborhood variability would have been expected to have appeared in the results of the GWLR variable coefficient maps because that is what the GWLR analysis method is intended to depict; the differences from one data point to the next. This was not the case in this research project as the coefficient maps only showed variability in Area 3 of the Combined Areas dataset used to make the predictions. Areas 1 and 2 showed little to no variability among the data. This was likely due to the design of this research project in both the sampling of the observed code violation data and in the bandwidth selection method. A true random sample of data across the entire city would have likely showed the variability in the data more accurately because there would have been data points that encompassed more of the distinct neighborhoods within the city. Furthermore, determining a bandwidth measure that encompassed only the subject SFR and properties immediately adjacent to it would have likely depicted more of the variability in the data within each study area. The optimum bandwidth selection tool in the GWR4 software was not able to find the spatial variation that is likely present in this data. Again, a true random sample of data points encompassing the entire city in conjunction with a more precise bandwidth selection method would likely give a large boost to the predictive capabilities of the GWLR analysis.

Research question #1 of this project was, "can certain property attributes predict the occurrence of code enforcement violations?" The answer to this is complicated and varies spatially, but generally the answer is yes. Some variables, such as number of stories or days to

previous violation, were always significant predictors. Others variables, such as structure age or corporation owned yes/no, were rarely found to be significant predictors, if at all. The general results of this analysis show there are in fact certain property characteristics that can predict the occurrence of housing code violations to some degree. However, the accuracy of these predictor variables is not strong enough to initiate code enforcement administrative action on because they do not account for all variations in the City's data and should be considered with great caution.

5.1.2. Spatial Findings

The geographically weighted logistic regression (GWLR) models showed that there was almost no spatial variability among the dependent variable and the independent variables. This was clear in the optimum bandwidth selection tool in the GWR4 program. This tool selects the bandwidth measure where the spatial variability is most pronounced in the data, which is not un-like the incremental spatial autocorrelation tool found in ArcGIS. In all of the random samples of the data, the optimum bandwidth was equal to the total size of the sample at 500. In Area 1, the bandwidth was 488 out of 581 homes, 216 out of 316 for Area 2, 954 out of 1,301 for Area 3, and 897 out of 2,198 for the combined areas. These large bandwidth sizes indicate that there is very little that can be explained geographically in the model. The initial thought in using GWLR was that houses that had code violations would have a "spill-over" effect on to nearby homes, similar to the effects of crime discussed in the writings of Block and Block (1995), Speer et al. (1998), Britt et al. (2005), and Toomey et al. (2012) as mentioned in Chapter 3. According to the GWR4 program calculations, this is not the case for code enforcement violations.

Existence of neighborhood variation was confirmed when the coefficient values of the combined areas, which was used to make the predictions, were mapped. For each variable coefficient, the results show that each neighborhood varies from the other two, but there is very

little variation inside the confines of each neighborhood area. The largest area, Area 3, was the only one that showed some degree or internal spatial variation in the coefficient values. Figure 12 below shows the days to violation coefficient values mapped across the three study areas. Notice that all points inside Area 1 fall between 0.5 and 1.2 standard deviations from the mean while all points in Area 2 are greater than -1.5 standard deviations from the mean. Most of Area 3 falls between -0.5 and 0.5 standard deviations with a few points falling in other ranges, though still within one standard deviation from the mean. So again, the model could not find any spatial variation among the single-family homes in the dataset. The remainder of the coefficient maps showing very similar outcomes can be found in Appendix E.

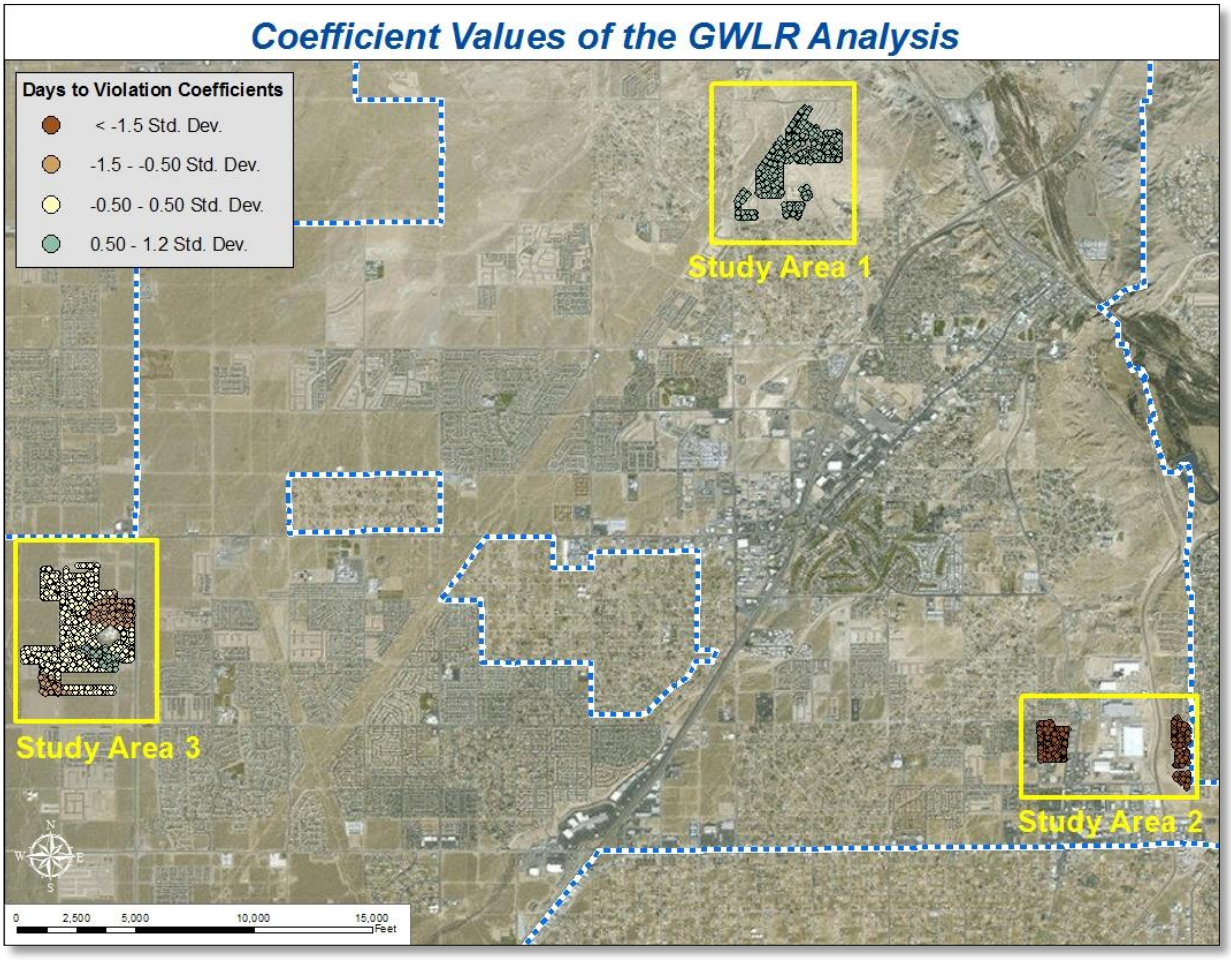


Figure 12: Map of the Days to Violation variable coefficient values from the GWLR analysis

The presence of variation in the neighborhoods is also present in the inability of the five random samples of the training data to make consistent predictions. As the results of the GWLR analysis show, two of the random samples of the data actually caused the model to decrease in prediction accuracy. In addition, the percent of deviance explained value was rather low compared to Area 1 and Area 2. Also, in the logistic regression model, the pseudo R squared values were not consistent and two of the samples had values that were at or below the acceptable threshold value of 0.20. So even though these samples are using the same dataset, there are differences in the data that are causing skewed results. This is likely because each random sample potentially has more samples from one neighborhood over another, and because the three study areas are so different, having a higher percentage of samples from one or the other is causing the data change. Section 5.3 explains how this phenomena can be reduced by employing a better data sampling method.

Research question 2 asked if the occurrence of a code enforcement violation is partially the result of the effect of neighboring properties, and the answer to this is no. The large optimum bandwidth values combined with the lack of spatial variability in the prediction model coefficient values shows that a spatial relationship between the dependent variable and the independent variables does not exist. However, as discussed earlier, using an alternative bandwidth selection method may reveal spatial relationships among neighboring properties. This is discussed in Section 5.3.

5.1.3. Predictions

The primary goal of this research project was the prediction of housing code violations throughout the City of Victorville. This was accomplished; however, the first thing to note here is the 50.3% accuracy when the model was validated using a proactive inspection area that was

collected after the initial data collection of this project. This is of course much lower than the 74.9% accuracy rate of the model training data. This result unfortunately suggests that the model was not a good predictor of housing code violations at the single-family home level. This is likely due to the weak relationships between the independent variables and the dependent variable found in the non-spatial logistic regression analysis of the odds ratios and their significance values.

However, this model has the potential to be able to predict housing code violations at the neighborhood level. A careful study of the four neighborhoods within the City of Victorville showed that homes in brand new housing tracts were largely predicted as having no violations. This is quite intuitive because these new homes are built to current code and have not had time to go unmaintained. In addition, neighborhoods that are regarded as "bad" areas of the City were predicted to have a higher density of possible violations, which is also intuitive and follows suit with the work of Meier (1983). In Meier (1983), he discussed how lower class neighborhoods have higher occurrence of code enforcement activity in Pasadena, CA. Areas in Victorville that are known to be lower class (in terms of quality) were predicted to have high occurrences of code violations. Though this could not be confirmed in the model validation process, the results of the predictions could be used in other methods to identify which neighborhoods are likely to have violations and which are not. This is discussed in Section 5.3.2.

This model was able to predict, to some degree, which houses in the City of Victorville currently have housing code violations, which answers research question 3. Of course, the previous discussions warrant extreme caution when acting upon these predictions due to low percentages of accuracy and weak relationships between the dependent and independent variables. This is also due to the fact that the accuracy was only tested on a small neighborhood

in the City, so neighborhoods with different characteristics could exhibit different degrees of prediction accuracy.

5.2 Relation to Previous Work

The results of this research project followed suit with many of the studies discussed in Chapter 2 of this document. The work of Des Rosiers et al. (2002) discussed how high quality landscaping can increase the value of a property. For new housing tracts in the City, development code standards require that a high quality landscape be installed prior to the home being permitted for occupancy by building inspectors. The model of this project predicted that these new homes are not likely to have code violations, and a contributing factor to this is a quality landscape, which leads to a high initial property value. Also, Luttik (2000) discussed how greenspace has a positive effect on surrounding property value. The model predicted that the homes immediately adjacent to the City's golf course had a low likelihood of housing code violations, which could be the result of higher home values according to Luttik (2000) and Meier (1983). Again Meier (1983) suggested that high property values lead to low code enforcement activity because residents are able to maintain their properties on their own.

This research project is also a contribution to the GIS based approaches to crime prediction. The binary dependent variable of code enforcement violation present yes/no could easily be changed to burglary present yes/no (Antolos et al. 2013) or simply crime present yes/no. The work of Murray et al. (2001) discussed many examples of how crime can be predicted using various spatial and statistical techniques such as Moran's I analysis or LISA analysis. Now, a logistic regression model could easily be included in this list because it is also employing methodology using spatial autocorrelation in the form of the GWLR technique. Crime was also modeled using logistic regression and GWR in the work of Wheeler and Waller

(2009), and Antolos et al. (2013). Wheeler and Waller (2009) were showing how GWR techniques can be improved on by using Bayesian regression and Antolos et al. (2013) used logistic regression to model the occurrence of burglary. This research project used ideas from both of these studies where the logistic regression method was initially used and was then improved upon by using a more refined regression technique, GWR in this case. Though the results of this research project did not coincide with the work of Antolos et al. (2013), the results do show that it is possible to model code violations in a manner similar to crime modeling.

This research project also adds another field that can be studied with logistic regression modeling. Section 2.3 of Chapter 2 outlined several studies where logistic regression was employed, including wildfire ignition (Perestrello de Vasconcelos et al. 2001), groundwater spring potential (Ozdemir 2011), and landslide susceptibility (Kundu et al. 2013). In each of these previous studies, the logistic regression technique was used to train a dataset to make predictions over a study area; this is exactly what was done in this research project. These previous studies were conducted for phenomena that occurred in the physical sciences. This research project adds a social science aspect to the realm of possible subjects that can be studied using the logistic regression technique.

Section 2.4 of Chapter 2 outlined several studies that employed geographically weighted regression techniques to make predictions of spatially varying phenomena, and some of them even went so far as to compare the GWR technique to the logistic regression technique. In Erener, Sebnem, and Düzgün (2010) and in Saefuddin, Setiabudi, and Fitrianto (2012), the researchers were able to improve the predictive capability of non-spatial regression models with the use of GWR. In this research project, it was found that there were very slight increases in the predictive accuracy of the GWLR model over its corresponding logistic regression model.

Future work on code enforcement prediction using the same methodology but stronger predictive variables could produce results that align with Erener, Sebnem, and Düzgün (2010) and Saefuddin, Setiabudi, and Fitrianto (2012) in terms of model improvement.

Finally, this research project also contributes to the work of Wu and Zhang (2013), Martinez-Fernandez, Chuvieco, and Koutsias (2013), and Rodrigues, de la Riva, and Fotheringham (2014) in that it gives another example where logistic and geographic regression modeling can be successful in making predictions over space. This research project was able to combine the methodologies in the aforementioned studies to produce a model that was able to make predictions with some degree of success. Even though these predictions were only 50% accurate at the parcel level when compared to a small sampling of observed code violations used for validation, it appeared to be successful in identifying neighborhood level clusters of violations that could potentially be addressed by code enforcement staff in the proactive inspection program. However, these neighborhood level clusters of violations were not field verified in this research project, and further investigation into the validity of identifying neighborhoods using parcel level predictions within logistic regression is necessary before this claim can be fully supported.

5.3 Limitations and Future Work

There were several observed limitations in this study. The results all indicate that there are likely several key variables that were not accounted for in this regression model. Also, the results show that there is little to no spatial variation between the dependent variable and its explanatory variables as indicated in the results of GWR4. This section addresses these limitations and what work can be done to improve these results.

5.3.1. Major Limitations

The most noticeable issue in this research project was the low pseudo R squared values in the non-spatial models and the low percentage of deviance explained values in the spatial models. As explained earlier, these values mean that there is much more to the story when it comes to predicting housing code violations in a city. This was one of the key contributing factors to the low observed accuracy of the predictions. The variables that were used in this research project were mostly related to the physical condition of the house, such as floor area, structure age, and assessed value. These variables did not include information on the neighborhood that these homes were in. Also, there were not enough variables that explained demographics or socio-economic condition, which play very important roles in determining the quality of a neighborhood and whether or not there are probable housing code violations. Furthermore, variables did not explain enough about the residents of each SFR, which will likely be a difficult explanatory variable to produce at the parcel level due to laws limiting data sharing from government agencies, including the US Census Bureau. Interpolations of US Census tract or block level data to arrive at parcel level data would also introduce the modifiable areal unit problem into the model. Also, conducting the analysis at the Census tract or block level in order to capture more demographic and economic data for the analysis would further emphasize the modifiable areal unit problem.

Another key limitation of this research project was the lack of significant relationships as explained by the odds ratio of the SPSS logistic regression outputs. As explained earlier, even though variables were determined to be statistically significant predictors, there were not many statistically significant relationships between the dependent variable and the explanatory variables. Many of the coefficient values were very close to zero and their converted odds ratios were extremely close or at 1, meaning that there was no significant relationship present. This

diminishes that overall quality of the model, which is a vital part of making accurate predictions.

Limitations were also noticed in the collection of the data. Code enforcement officers exercise a great deal of discretion when determining if a home has a housing code violation. Each officer's interpretation of the code could be different from the others, meaning that there could be data discrepancies based on which officers performed the initial proactive inspections. This actually relates to the work of Ross (1996) and Burby et al. (2000) where they discussed government discretion and how it varies between government officials, interpretation of codes, and how this can cause violations to be overlooked. This potentially could create data quality issues with the model. This was not specifically analyzed in this research project, but it would be worth studying.

Having three sample areas that were miles apart was also a limitation in this analysis. Because of this, it was difficult to produce a model that encompassed more diversity among homes and neighborhoods so that these factors could be accounted for in the predictions. It would have been more accurate to take a random sample of the City in its entirety rather than random samples of data within the three study areas. This would have been able to capture a more accurate depiction of the variability in the City's SFRs. The City contains many diverse neighborhoods and having data from only three of these neighborhoods produced low accuracy when tested against a neighborhood that had different characteristics than the sample areas. The fact that the City has so much variability in its housing and that codes can be interpreted differently by different officers or inspectors makes predicting violations a very difficult task, and the results of this research project are a good indicator of that.

Finally, training the model using clusters of data points instead of across the entire study area could potentially bias the prediction model. For example, if there are high code violation

counts that come from the clusters of training data, the model will have a hard time differentiating data points where there is a low likelihood of a violation because the model does not have any examples of this. It would be the same kind of bias if the training data had a low occurrence of code violations as it would not be able to pin-point SFRs with a high likelihood of a violation with any ease. The fact that logistic regression uses a binary dependent variable makes minimizing this model bias difficult because the logistic model cannot determine the degree or severity of the violation. This bias could potentially be reduced by either having a good sampling of data points that encompasses neighborhoods with high violations and neighborhoods with low violations, or by making the dependent variable nominal or ordinal so the model can have other training information to be used in predictions. Also, changing the dependent variable to nominal or ordinal would require a different regression technique, meaning that the binary logistic technique of this research project could not be used.

5.3.2. Future Work

There is a great deal of room for future work with this research project. The first area would be to find better prediction variables. As mentioned earlier, variables that incorporate demographic or socio-economic factors such as household income or a normalization of household income against the assessed property value would be useful. As Meier (1983) discovered in his study, areas with higher household income levels saw a much lower need for code enforcement action because property owners were financially able to maintain their properties as required by city regulations. Another variable that should be incorporated would be some sort of neighborhood variable. There needs to be something that can describe the state of a neighborhood to make stronger predictions and account for differences among neighborhoods. These variables could consider crime rates, dwelling unit density, ethnic groups, age of residents,

single parent households, or educational achievement.

Other methods of variable collection could also be employed. Remote sensing could potentially be used to capture SFRs where the front yard has gone from green grass to brown or dead grass using infrared sensors and image classification software. This would show SFRs that violate the live landscaping requirement of the City. Aerial photography could also be used to show SFRs with the presence of inoperable vehicles. For example, if an aerial photograph shows a vehicle in the driveway or on the street in a February flight, and that vehicle appears in a July flight and has not moved, there is a good chance that the vehicle does not run; Code says vehicles that do not run must not be visible from the street. Aerial photography could also be used to locate SFRs that have trash and debris that must be abated using basic visual scans of photographed areas. Using remote sensing would be a time consuming and expensive process, but it could yield information that is vital to predicting if other SFRs are likely to have the same type of code violations.

Hot spot analysis of the code violation predictions could be useful in strengthening the claim that the model created in this research project was potentially able to identify neighborhoods where there is a higher likelihood of violations. So instead of visually scanning each neighborhood in the prediction data to determine if there is a high or low occurrence of violations, the hot spot analysis could identify these areas more quickly. Furthermore, the hot spot analysis could be used to identify neighborhoods with the highest intensity of violations, as well as neighborhoods with the lowest intensity. This would help City officials to allocate resources to the most intense hot spots first, followed by less intense hotspots. This could produce quick changes in the neighborhood characteristics of these areas, and it would show elected officials of the City that the proactive enforcement program does have some degree of

impact, hopefully in the positive direction.

Future work should also include more sound sampling methods to produce more consistent datasets. Having three clusters of data in three very different neighborhoods likely caused issues in the project. If data collection practices were more systematic, say by using only one officer to do all of the inspections, and if more data was collected from different neighborhoods, such as the half acre lot neighborhoods which are much lower in dwelling density, the model could potentially recognize more of the neighborhood variability in the data and make stronger predictions. Also, adding more areas to verify the accuracy of the model should increase the 50.3% observed accuracy number.

5.3.3. Conclusions

Thus, it can be seen that even though this project was not able to produce highly accurate predictions of housing code violations, there were achievements in determining how well selected variables performed in making predictions and in determining how much of a role basic geography principles played in this phenomena. Furthermore, the prediction model generated using the regression techniques could potentially be used to identify neighborhoods that have a high likelihood of violations and others that have a low likelihood of violations. This is based on observed knowledge of these neighborhoods and how the predictions seem to correlate to the neighborhood characteristics. Also, identifying which variables expressed multicollinearity with other variables will aid in the replication of this research project because there will not be such a strong effort on data collection.

In replication of this research project by other cities, the advice would be to obtain stronger predictor variables, collect a random sample data from the entire city instead of three clusters of SFRs in order to better explain the role of geography, and to use more validation

areas.. Stronger predictor variables could make the relationship between the dependent and independent variables stronger. Also, collecting a random sample of data points across the entire city area would make the kernel bandwidth in the GWLR analysis more equipped to depict the spatial relationships in the data. Finally, providing more validation areas could potentially increase the model's prediction accuracy to a point where it could be trusted. Doing so could make this prediction model more accurate in order to one day be able to make code enforcement departments in any city more effective.

REFERENCES

Anderson, J. A. 1982. Logistic regression. *Handbook of Statistics. North-Holland, New York*: 169-191.

Anderson, Linda M., and H. Ken Cordell. 1988. Influence of trees on residential property values in Athens, Georgia (USA): a survey based on actual sales prices. *Landscape and Urban Planning* 15 (1): 153-164.

Antolos, D., D. Liu, A. Ludu, and D. Vincenzi. 2013. Burglary crime analysis using logistic regression. *Human Interface and the Management of Information. Information and Interaction for Learning, Culture, Collaboration and Business, Lecture Notes in Computer Science* 8018: 549-558.

Atkinson, P. M., S. E. German, D. A. Sear, and M. J. Clark. 2003. Exploring the relations between riverbank erosion and geomorphological controls using geographically weighted logistic regression. *Geographical Analysis* 35 (1): 58-82.

Block, R. L., and C. R. Block. 1995. Space, place and crime: hot spot areas and hot places of liquor-related crime. *Crime and Place* 4 (2): 145-184.

Britt, H. R., B. P. Carlin, T. L. Toomey, and A. C. Wagenaar. 2005. Neighborhood level spatial analysis of the relationship between alcohol outlet density and criminal violence. *Environmental and Ecological Statistics* 12 (4): 411-426.

Burby, R. J., P. J. May, and R. C. Paterson. 1998. Improving compliance with regulations: choices and outcomes for local government. *Journal of the American planning association* 64 (3): 324-334.

Carlton, R. E. 1965. Enforcement of municipal housing codes. *Harvard law review* 78 (4): 801-860.

Chainey, S., L. Tompson, and S. Uhlig. 2008. The utility of hot spot mapping for predicting spatial patterns of crime. *Security Journal* (19) 1-2: 4-28

Clark, W. A., and P. L. Hosking. 1986. Statistical methods for geographers. *Wiley, New York*.

Conway, D., C. Q. Li, J. Wolch, C. Kahle, and M. Jerrett. 2008. A spatial autocorrelation approach for examining the effects of urban greenspace on residential property values. *Journal of Real Estate, Finance, and Economics* (2010) 41: 150-169.

Des Rosiers, F., M. Theriault, Y. Kestens, and P. Villenevue. 2002. Landscaping and housing values: an empirical investigation. *The Journal of Real Estate Research* 23 (1): 139-161.

Dombrow, J., and M. Rodriguez. 2000. The market value of mature trees in single-family housing markets. *The Appraisal Journal* 68 (1): 39-43.

Erener, A., H. Sebnem, B. Düzgün. 2010. Improvement of statistical landslide susceptibility mapping by using spatial and global regression methods in the case of More and Romsdal (Norway). *Landslides* 7: 55-68.

Gerber, M. S. 2014. Predicting crime using Twitter and kernel density estimation. *Decision Support Systems* (61): 115-125.

Gibbons, S. 2004. The costs of urban property crime. *The Economic Journal* 114 (499): F441-F463.

Gribetz, J., and F. P. Grad. 1966. Housing code enforcement: sanctions and remedies. *Columbia law review* 66 (7): 1254-1290.

Grubesic, T. H., and A. T. Murray. 2001. Detecting hot spots using cluster analysis and GIS. *Proceedings from the Fifth Annual International Crime Mapping Research Conference* 26.

Kundu, S., A. K. Saha, D. C. Sharma, and C. C. Pant. 2013. Remote sensing and GIS based landslide susceptibility assessment using binary logistic regression model: a case study in Ganeshganga watershed, Himalayas. *Journal of the Indian Society of Remote Sensing* 41 (3): 697-709.

Lee, S., and t. Sambath. 2006. Landslide susceptibility mapping in the Damrei Romel area, Cambodia using frequency ratio and logistic regression models. *Environmental Geology* 50: 847-855.

Linden, L., and J. E. Rockoff. 2008. Estimates of the impact of crime risk on property values from Megan's laws. *The American Economic Review*: 1103-1127.

Liu, H., and D. E. Brown. 2003. Criminal incident prediction using a point-pattern-based density model. *International Journal of Forecasting* 19: 603-622.

Luo, J., N. K. Kanala. 2008. Modeling urban growth with geographically weighted multinomial logistic regression. *Geoinformatics* 7144: 7144M-1 – M-11.

Luttik, J. 2000. The value of trees, water and open space as reflected by house prices in the Netherlands. *Landscape and Urban Planning* 48 (3): 161-167.

Martinez-Fernandez, J., E. Chuvieco, and N. Koutsias. 2013. Modelling long-term fire occurrence factors in Spain by accounting for local variations with geographically weighted regression. *Natural Hazards Earth System Science* 13: 311-327.

Meier, R. B. 1983. Code enforcement and housing quality revisited: the turnover case. *Urban Affairs Quarterly* 19 (2): 255-273.

Miller Jr., L. Charles. 1973. The economics of housing code enforcement. *Land Economics* 49 (1): 92-96.

Morckel, V. 2014. Predicting abandoned housing: does the operational definition of abandonment matter? *Community Development* 45 (2): 121-133.

Murray, A. T., I. McGuffog, J. S. Western, and P. Mullins. 2001. Exploratory spatial data analysis techniques for examining urban crime. *The British Journal of Criminology* (41): 309-329.

Nakaya, T., K. Yano. 2010. Visualizing crime clusters in a space-time cube: an exploratory data-analysis approach using space-time kernel density estimation and scan statistics. *Transactions in GIS* 14 (3): 223-239.

Nakaya, T. Geographically Weighted Regression (version 4.0). Windows. Tempe, AZ: Arizona State University GeoDa Center for Geospatial Analysis and Computation, 2009.

Osgood, D. W., L. L. Finken, and B. J. McMorris. 2002. Analyzing multiple-item measures of crime and deviance II: Tobit regression analysis of transformed scores. *Journal of Quantitative Criminology* 18 (4): 319-347.

Ozdemir, A. 2011. Using a binary logistic regression method and GIS for evaluating and mapping the groundwater spring potential in the Sultan Mountains (Aksehir, Turkey). *Journal of Hydrology* 405: 123-136.

Perestrello, M. J., S. Silva, M. Tomé, M. Alvim, and J. M. Cardoso Pereira. 2001. Spatial prediction of fire ignition probabilities: comparing logistic regression and neural networks. *Photogrammetric Engineering & Remote Sensing* 67 (1): 73-81.

Rodrigues, M., J. de la Riva, and S. Fotheringham. 2014. Modeling the spatial variation of the explanatory factors of human-caused wildfires in Spain using geographically weighted logistic regression. *Applied Geography* 48: 52-63.

Ross, H. L. 1996. Housing code enforcement and urban decline. *Journal of Affordable Housing & Community Development Law* 6 (1): 29-46.

Saefuddin, A., N. A. Setiabudi, and A. Fitrianto. 2012. On comparison between logistic regression and geographically weighted logistic regression: with application to Indonesian poverty data. *World Applied Sciences Journal* 19 (2): 205-210.

Simon, Steve. 2013. Calculating predicted probabilities from a logistic regression model. *The Monthly Mean*. July 31. Accessed November 27, 2015. http://www.pmean.com/13/predicted.html

Speer, P. W., D. M. Gorman, E. W. Labouvie, and M. J. Ontkush. 1998. Violent crime and alcohol availability: relationships in an urban community. *Journal of Public Health Policy* 19 (3): 303-318

Teater, B. A. 2011. A qualitative evaluation of the section 8 housing choice voucher program. *Qualitative Social Work* 10 (4): 503-519.

Tobler, W. R. 1970. A computer movie simulating urban growth in the Detroit region. *Economic Geography* 46: 234-240.

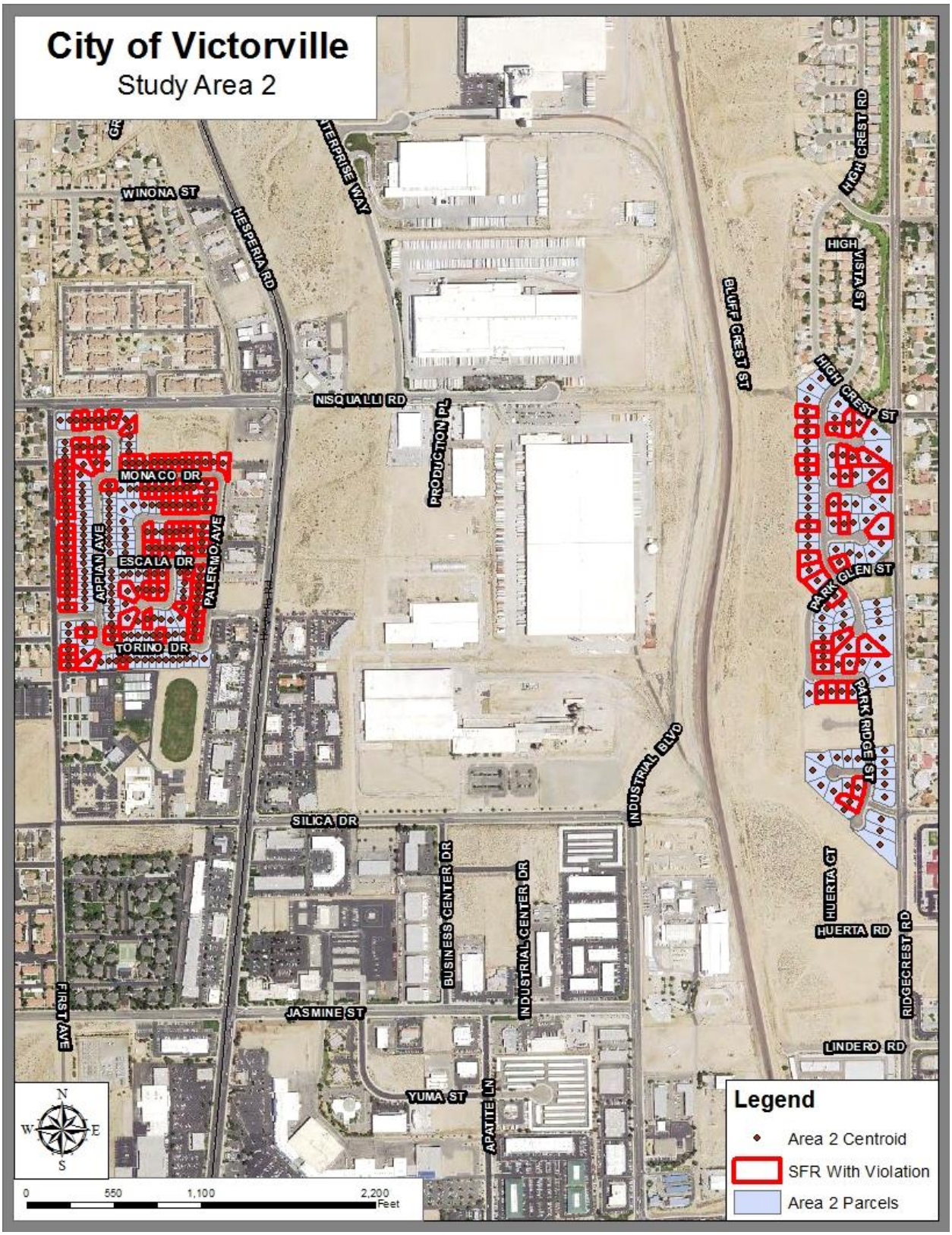
Toomey, T. L., D. J. Erickson, B. P. Carlin, K. M. Lenk, H. S. Quick, A. M. Jones, and E. M. Harwood. 2012. The association between density of alcohol establishments and violent crime within urban neighborhoods. *Alcoholism: Clinical and Experimental Research* 36 (8): 1468-1473.

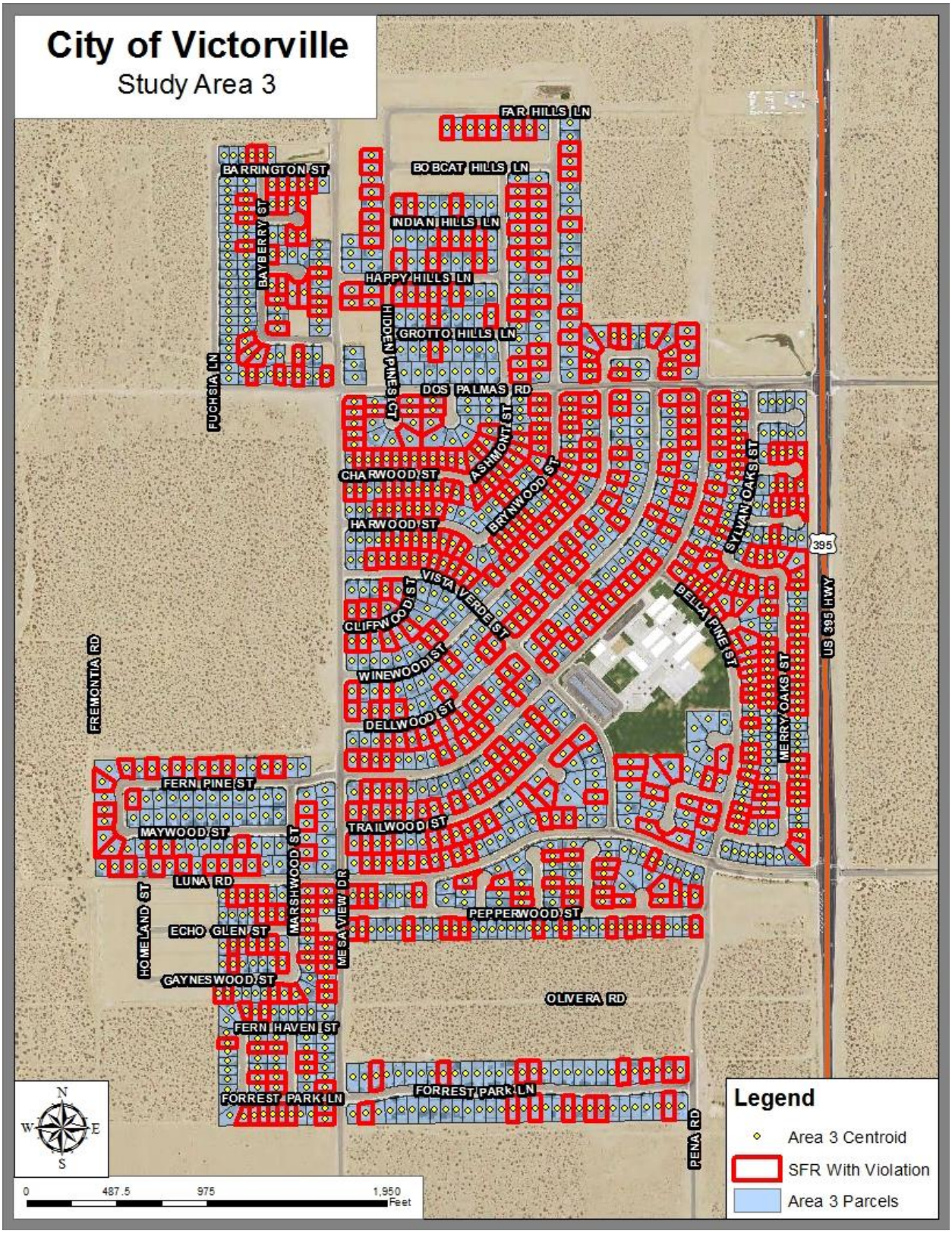
Wheeler, D.C, and L. A. Waller. 2008. Comparing spatially varying coefficient models: a case study examining violent crime rates and their relationships to alcohol outlets and illegal drug arrests. *Journal of Geographic Systems* (2009) 11 (1): 1-22.

Wu, W., L. Zhang. 2013. Comparison of spatial and non-spatial logistic regression models for modeling the occurrence of cloud cover in north-eastern Puerto Rico. *Applied Geography* 37: 52-62.

Xue, Y., and D. E. Brown. 2006. Spatial analysis with preference specification of latent decision makers for criminal event prediction. *Decision support systems* 41: 560-573.

Appendix A: Area 2 and Area 3 Maps




City of Victorville
Study Area 3
FAR HILLS LN
BARRINGTON ST
BOBCAT HILLS LN
INDIAN HILLS LN
BAYBERRY ST
HAPPY HILLS LN
HIDDEN PINES CT
GROTTO HILLS LN
FUCHSIA LN
DOS PALMAS RD
ASHMONT ST
CHARWOOD ST
BRYNWOOD ST
HARWOOD ST
SYLVAN OAKS ST
395
VISTA VERDE ST
CLIFFWOOD ST
BELLA PINE ST
US 395 HWY
WINEWOOD ST
FREMONTIA RD
DELLWOOD ST
MERRY OAKS ST
FERN PINE ST
MAYWOOD ST
TRAILWOOD ST
MARSHWOOD ST
LUNA RD
MESA VIEW DR
HOMELAND ST
PEPPERWOOD ST
ECHO GLEN ST
GAYNESWOOD ST
OLIVERA RD
FERN HAVEN ST
FORREST PARK LN
FORREST PARK LN
PENA RD
N
W
E
S
0
487.5
975
1,950
Feet
Legend
Area 3 Centroid
SFR With Violation
Area 3 Parcels

Appendix B Example SPSS Output

Dependent Variable Encoding

Original Value	Internal Value
NO	0
YES	1

Categorical Variables Codings

		Frequency	Parameter coding (1)
NOSTORY	1	170	1.000
	2	411	.000

Block 0: Beginning Block

Classification Table[a,b]

Observed			Predicted		
			PROACTIVE		Percentage Correct
			NO	YES	
Step 0	PROACTIVE	NO	438	0	100.0
		YES	143	0	.0
	Overall Percentage				75.4

a. Constant is included in the model.

b. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 0	Constant	-1.119	.096	135.078	1	.000	.326

Variables not in the Equation[a]

			Score	df	Sig.
Step 0	Variables	LOTSQFT	5.252	1	.022
		FLOOR_AREA	39.343	1	.000
		NOSTORY(1)	14.777	1	.000
		LENGTH_OWN	6.810	1	.009
		DAYS_TO_VI	85.649	1	.000
		CASE_COUNT	73.827	1	.000
		NETWORK_DI	26.390	1	.000
		NEARBY_CAS	35.841	1	.000

a. Residual Chi-Squares are not computed because of redundancies.

Block 1: Method = Enter

Omnibus Tests of Model Coefficients

		Chi-square	df	Sig.
Step 1	Step	154.096	8	.000
	Block	154.096	8	.000
	Model	154.096	8	.000

Model Summary

Step	-2 Log likelihood	Cox & Snell R Square	Nagelkerke R Square
1	494.347[a]	.233	.346

a. Estimation terminated at iteration number 5 because parameter estimates changed by less than .001.

Classification Table[a]

			Predicted		
			PROACTIVE		Percentage Correct
	Observed		NO	YES	
Step 1	PROACTIVE	NO	410	28	93.6
		YES	63	80	55.9
	Overall Percentage				84.3

a. The cut value is .500

Variables in the Equation

		B	S.E.	Wald	df	Sig.	Exp(B)
Step 1[a]	LOTSQFT	.000	.000	.931	1	.335	1.000
	FLOOR_AREA	-3.978	1.772	5.040	1	.025	.019
	NOSTORY(1)	-.140	.326	.185	1	.667	.869
	LENGTH_OWN	.000	.000	4.281	1	.039	1.000
	DAYS_TO_VI	.000	.000	22.152	1	.000	1.000
	CASE_COUNT	.147	.115	1.645	1	.200	1.158
	NETWORK_DI	.000	.000	20.211	1	.000	1.000
	NEARBY_CAS	.059	.033	3.304	1	.069	1.061
	Constant	2.396	1.242	3.718	1	.054	10.976

Variables in the Equation

		95% C.I.for EXP(B)	
		Lower	Upper
Step 1[a]	LOTSQFT	1.000	1.000
	FLOOR_AREA	.001	.604
	NOSTORY(1)	.459	1.647
	LENGTH_OWN	1.000	1.000
	DAYS_TO_VI	.999	1.000
	CASE_COUNT	.925	1.450
	NETWORK_DI	1.000	1.000
	NEARBY_CAS	.995	1.132
	Constant		

a. Variable(s) entered on step 1: LOTSQFT, FLOOR_AREA, NOSTORY, LENGTH_OWN, DAYS_TO_VI, CASE_COUNT, NETWORK_DI, NEARBY_CAS.

Appendix C Logistic Regression Results Full Summary Table

	Multicollinearity		Predictors		Coefficients & Relationships		
Variable Name and (Alias)	Initial VIF Value	Final VIF Value	Predictor Sig.	Keep In Model?	Variable Coefficient	Variable Odds Ratio - Exp(B)	Odds Ratio Sig.
AREA 1							
Lot Size (LOTSQFT)	4.116	1.520	0.022	YES	0.000	1.000	0.335
Floor Area (FLOOR_AREA)	1.889	1.718	0.000	YES	-3.978	0.019	0.025
Number of Stories (NOSTORY)	2.842	1.475	0.000	YES	-0.140	0.869	0.667
Length of Ownership (LENGTH_OWN)	1.184	1.088	0.009	YES	0.000	0.039	1.000
Structure Age (STRUCT_AGE)	1.047	1.020	0.208	NO			
Total Assessed Value (TOTAL_VALU)	7.646	REMOVED					
Value per Sq Ft (VALUE_PSF)	4.186	1.215	0.698	YES	0.001	1.001	0.928
Code Case Yes/No (CODECASE)	5.955	REMOVED					
Days to Previous Violation (DAYS_TO_VI)	7.812	2.314	0.000	YES	0.000	1.000	0.000
Number of Previous Cases (CASE_COUNT)	2.553	2.443	0.000	YES	0.147	0.200	1.158
Tax Defaulted Yes/No (TAX_STATUS)	1.033	1.019	0.208	NO			
Occupancy (OCCUPANCY)	1.152	1.035	0.934	NO			
Cartesian Distance to Liquor Store (NEAR_DIST)	5.772	REMOVED					
Travel Distance to Liquor Store (NETWORK_DIST)	1.171	1.106	0.000	YES	0.000	1.000	0.000
Number of Nearby Cases (NEARBY_CASE)	3.229	1.966	0.000	YES	0.059	1.061	0.069
Corporate Ownership (CORP_OWNED)	1.172	1.154	0.504	NO			
Constant					2.396	10.976	0.054
AREA 2							
Lot Size (LOTSQFT)	4.116	REMOVED					
Floor Area (FLOOR_AREA)	1.889	1.718	0.265	NO			
Number of Stories (NOSTORY)	2.842	1.475	0.820	NO			

	Multicollinearity		Predictors		Coefficients & Relationships		
Variable Name and (Alias)	Initial VIF Value	Final VIF Value	Predictor Sig.	Keep In Model?	Variable Coefficient	Variable Odds Ratio - Exp(B)	Odds Ratio Sig.
Length of Ownership (LENGTH_OWN)	1.184	1.088	0.811	NO			
Structure Age (STRUCT_AGE)	1.047	1.020	0.640	NO			
Total Assessed Value (TOTAL_VALU)	7.646	REMOVED					
Value per Sq Ft (VALUE_PSF)	4.186	1.215	0.048	YES	0.001	1.001	0.928
Code Case Yes/No (CODECASE)	5.955	REMOVED					
Days to Previous Violation (DAYS_TO_VI)	7.812	2.314	0.000	YES	-0.001	0.999	0.000
Number of Previous Cases (CASE_COUNT)	2.553	2.443	0.000	YES	0.067	1.069	0.695
Tax Defaulted Yes/No (TAX_STATUS)	1.033	1.019	0.643	NO			
Occupancy (OCCUPANCY)	1.152	1.035	0.206	NO			
Cartesian Distance to Liquor Store (NEAR_DIST)	5.772	REMOVED					
Travel Distance to Liquor Store (NETWORK_DIST)	1.171	1.106	0.845	NO			
Number of Nearby Cases (NEARBY_CASE)	3.229	1.966	0.000	YES	0.001	1.001	0.973
Corporate Ownership (CORP_OWNED)	1.172	1.154	0.616	NO			
Constant					1.261	3.527	0.154
AREA 3							
Lot Size (LOTSQFT)	5.881	REMOVED					
Floor Area (FLOOR_AREA)	8.405	1.726	0.247	NO			
Number of Stories (NOSTORY)	2.656	1.752	0.156	NO			
Length of Ownership (LENGTH_OWN)	1.202	1.170	0.026	YES	0.000	1.000	0.109
Structure Age (STRUCT_AGE)	1.006	1.004	0.775	NO			
Total Assessed Value (TOTAL_VALU)	7.826	REMOVED					
Value per Sq Ft (VALUE_PSF)	6.509	1.345	0.452	NO			

	Multicollinearity		Predictors		Coefficients & Relationships		
Variable Name and (Alias)	Initial VIF Value	Final VIF Value	Predictor Sig.	Keep In Model?	Variable Coefficient	Variable Odds Ratio - Exp(B)	Odds Ratio Sig.
Code Case Yes/No (CODECASE)	9.396	REMOVED					
Days to Previous Violation (DAYS_TO_VI)	2.843	2.843	0.000	YES	-0.001	0.999	0.000
Number of Previous Cases (CASE_COUNT)	2.823	1.115	0.000	YES	-0.250	0.779	0.033
Tax Defaulted Yes/No (TAX_STATUS)	1.005	1.004	0.357	NO			
Occupancy (OCCUPANCY)	1.161	1.149	0.003	YES	0.071	1.074	0.668
Cartesian Distance to Liquor Store (NEAR_DIST)	1.191	1.191	0.015	YES	0.000	1.000	0.201
Travel Distance to Liquor Store (NETWORK_DIST)	1.184	1.181	0.003	YES	0.000	1.000	0.015
Number of Nearby Cases (NEARBY_CASE)	1.120	1.082	0.000	YES	0.034	1.034	0.140
Corporate Ownership (CORP_OWNED)	1.110	1.114	0.278	NO			
Constant					2.114	8.284	0.000
RANDOM SAMPLE 1							
Lot Size (LOTSQFT)	6.066	REMOVED					
Floor Area (FLOOR_AREA)	8.364	1.541	0.434	NO			
Number of Stories (NOSTORY)	2.178	1.552	0.071	NO			
Length of Ownership (LENGTH_OWN)	1.117	1.111	0.314	NO			
Structure Age (STRUCT_AGE)	1.033	1.031	0.172	NO			
Total Assessed Value (TOTAL_VALU)	7.542	REMOVED					
Value per Sq Ft (VALUE_PSF)	5.998	1.262	0.351	NO			
Code Case Yes/No (CODECASE)	9.091	REMOVED					
Days to Previous Violation (DAYS_TO_VI)	2.736	2.732	0.000	YES	0.000	1.000	0.000
Number of Previous Cases (CASE_COUNT)	2.820	2.803	0.000	YES	-0.016	0.984	0.919
Tax Defaulted Yes/No (TAX_STATUS)	1.034	1.033	0.927	NO			

	Multicollinearity		Predictors		Coefficients & Relationships		
Variable Name and (Alias)	Initial VIF Value	Final VIF Value	Predictor Sig.	Keep In Model?	Variable Coefficient	Variable Odds Ratio - Exp(B)	Odds Ratio Sig.
Occupancy (OCCUPANCY)	1.140	1.130	0.549	NO			
Cartesian Distance to Liquor Store (NEAR_DIST)	1.746	1.738	0.063	NO			
Travel Distance to Liquor Store (NETWORK_DIST)	1.694	1.694	0.095	NO			
Number of Nearby Cases (NEARBY_CASE)	1.338	1.198	0.000	YES	0.060	1.062	0.035
Corporate Ownership (CORP_OWNED)	1.088	1.088	0.551	NO			
Constant					0.351	1.421	0.369
RANDOM SAMPLE 2							
Lot Size (LOTSQFT)	7.635	REMOVED					
Floor Area (FLOOR_AREA)	10.123	1.530	0.149	NO			
Number of Stories (NOSTORY)	2.308	1.581	0.001	YES	0.839	2.315	0.000
Length of Ownership (LENGTH_OWN)	1.136	1.127	0.135	NO			
Structure Age (STRUCT_AGE)	1.024	1.020	0.121	NO			
Total Assessed Value (TOTAL_VALU)	8.783	REMOVED					
Value per Sq Ft (VALUE_PSF)	6.818	1.291	0.564	NO			
Code Case Yes/No (CODECASE)	9.475	REMOVED					
Days to Previous Violation (DAYS_TO_VI)	10.677	2.253	0.000	YES	-0.001	0.999	0.000
Number of Previous Cases (CASE_COUNT)	2.353	2.344	0.000	YES	-0.168	0.846	0.210
Tax Defaulted Yes/No (TAX_STATUS)	1.054	1.046	0.303	NO			
Occupancy (OCCUPANCY)	1.140	1.071	0.892	NO			
Cartesian Distance to Liquor Store (NEAR_DIST)	1.821	1.804	0.048	YES	0.000	1.000	0.926
Travel Distance to Liquor Store (NETWORK_DIST)	1.659	1.633	0.010	YES	0.000	1.000	0.058
Number of Nearby Cases (NEARBY_CASE)	1.456	1.302	0.000	YES	0.016	1.016	0.637

	Multicollinearity		Predictors		Coefficients & Relationships		
Variable Name and (Alias)	Initial VIF Value	Final VIF Value	Predictor Sig.	Keep In Model?	Variable Coefficient	Variable Odds Ratio - Exp(B)	Odds Ratio Sig.
Corporate Ownership (CORP_OWNED)	1.112	1.112	0.918	NO			
Constant					1.225	3.405	0.034
RANDOM SAMPLE 3							
Lot Size (LOTSQFT)	1.082	1.510	0.336	NO			
Floor Area (FLOOR_AREA)	1.807	2.170	0.001	YES	-2.158	0.116	0.094
Number of Stories (NOSTORY)	2.075	1.676	0.004	YES	0.399	1.491	0.112
Length of Ownership (LENGTH_OWN)	1.172	1.170	0.740	NO			
Structure Age (STRUCT_AGE)	1.044	1.037	0.033	YES	0.007	1.007	0.359
Total Assessed Value (TOTAL_VALU)	3.248	REMOVED					
Value per Sq Ft (VALUE_PSF)	2.805	1.311	0.796	NO			
Code Case Yes/No (CODECASE)	9.376	REMOVED					
Days to Previous Violation (DAYS_TO_VI)	10.516	2.092	0.000	YES	0.000	1.000	0.000
Number of Previous Cases (CASE_COUNT)	2.223	2.196	0.000	YES	-0.091	0.913	0.446
Tax Defaulted Yes/No (TAX_STATUS)	1.053	1.047	0.153	NO			
Occupancy (OCCUPANCY)	1.207	1.148	0.314	NO			
Cartesian Distance to Liquor Store (NEAR_DIST)	1.969	1.953	0.022	YES	0.000	1.000	0.685
Travel Distance to Liquor Store (NETWORK_DIST)	1.804	1.776	0.000	YES	0.000	1.000	0.009
Number of Nearby Cases (NEARBY_CASE)	1.386	1.293	0.001	YES	0.019	1.019	0.554
Corporate Ownership (CORP_OWNED)	1.085	1.075	0.136	NO			
Constant					2.119	8.324	0.011
RANDOM SAMPLE 4							
Lot Size (LOTSQFT)	7.351	REMOVED					
Floor Area (FLOOR_AREA)	8.822	1.557	0.163	NO			

	Multicollinearity		Predictors		Coefficients & Relationships		
Variable Name and (Alias)	Initial VIF Value	Final VIF Value	Predictor Sig.	Keep In Model?	Variable Coefficient	Variable Odds Ratio - Exp(B)	Odds Ratio Sig.
Number of Stories (NOSTORY)	2.082	1.573	0.032	YES	0.598	1.818	0.008
Length of Ownership (LENGTH_OWN)	1.226	1.209	0.504	NO			
Structure Age (STRUCT_AGE)	1.018	1.016	0.304	NO			
Total Assessed Value (TOTAL_VALU)	7.917	REMOVED					
Value per Sq Ft (VALUE_PSF)	5.786	1.395	0.025	YES	-0.017	0.983	0.005
Code Case Yes/No (CODECASE)	9.925	REMOVED					
Days to Previous Violation (DAYS_TO_VI)	11.466	2.510	0.000	YES	0.000	1.000	0.000
Number of Previous Cases (CASE_COUNT)	2.698	2.682	0.000	YES	0.012	1.012	0.938
Tax Defaulted Yes/No (TAX_STATUS)	1.052	1.035	0.284	NO			
Occupancy (OCCUPANCY)	1.233	1.164	0.001	YES	-0.168	0.845	0.510
Cartesian Distance to Liquor Store (NEAR_DIST)	1.955	1.952	0.096	NO			
Travel Distance to Liquor Store (NETWORK_DIST)	1.730	1.740	0.006	YES	0.000	1.000	0.013
Number of Nearby Cases (NEARBY_CASE)	1.498	1.329	0.000	YES	0.061	1.063	0.039
Corporate Ownership (CORP_OWNED)	1.103	1.105	0.879	NO			
Constant					2.111	8.254	0.004
RANDOM SAMPLE 5							
Lot Size (LOTSQFT)	6.725	REMOVED					
Floor Area (FLOOR_AREA)	9.282	1.547	0.016	YES	-0.479	0.620	0.727
Number of Stories (NOSTORY)	2.291	1.592	0.000	YES	0.602	1.826	0.025
Length of Ownership (LENGTH_OWN)	1.175	1.159	0.698	NO			
Structure Age (STRUCT_AGE)	1.023	1.021	0.173	NO			
Total Assessed Value (TOTAL_VALU)	7.295	REMOVED					

	Multicollinearity		Predictors		Coefficients & Relationships		
Variable Name and (Alias)	Initial VIF Value	Final VIF Value	Predictor Sig.	Keep In Model?	Variable Coefficient	Variable Odds Ratio - Exp(B)	Odds Ratio Sig.
Value per Sq Ft (VALUE_PSF)	5.160	1.250	0.573	NO			
Code Case Yes/No (CODECASE)	7.867	REMOVED					
Days to Previous Violation (DAYS_TO_VI)	9.722	2.598	0.000	YES	-0.001	0.999	0.000
Number of Previous Cases (CASE_COUNT)	2.613	2.595	0.000	YES	-0.158	0.854	0.320
Tax Defaulted Yes/No (TAX_STATUS)	1.020	1.018	0.423	NO			
Occupancy (OCCUPANCY)	1.167	1.110	0.003	YES	-0.170	0.844	0.518
Cartesian Distance to Liquor Store (NEAR_DIST)	1.888	1.873	0.010	YES	0.000	1.000	0.864
Travel Distance to Liquor Store (NETWORK_DIST)	1.802	1.771	0.001	YES	0.000	1.000	0.101
Number of Nearby Cases (NEARBY_CASE)	1.320	1.220	0.000	YES	0.055	1.056	0.086
Corporate Ownership (CORP_OWNED)	1.115	1.113	0.248	NO			
Constant					1.414	4.111	0.110
COMBINED AREAS							
Lot Size (LOTSQFT)	2.682	REMOVED					
Floor Area (FLOOR_AREA)	4.073	1.551	0.000	YES	-0.623	0.536	0.309
Number of Stories (NOSTORY)	2.144	1.580	0.000	YES	0.437	1.548	0.000
Length of Ownership (LENGTH_OWN)	1.157	1.139	0.503	NO			
Structure Age (STRUCT_AGE)	1.002	1.002	0.597	NO			
Total Assessed Value (TOTAL_VALU)	4.048	REMOVED					
Value per Sq Ft (VALUE_PSF)	3.286	1.288	0.444	NO			
Code Case Yes/No (CODECASE)	9.106	REMOVED					
Days to Previous Violation (DAYS_TO_VI)	10.761	2.410	0.000	YES	0.000	1.000	0.000
Number of Previous Cases (CASE_COUNT)	2.516	2.495	0.000	YES	-0.088	0.916	0.203

	Multicollinearity		Predictors		Coefficients & Relationships		
Variable Name and (Alias)	Initial VIF Value	Final VIF Value	Predictor Sig.	Keep In Model?	Variable Coefficient	Variable Odds Ratio - Exp(B)	Odds Ratio Sig.
Tax Defaulted Yes/No (TAX_STATUS)	1.010	1.009	0.341	NO			
Occupancy (OCCUPANCY)	1.173	1.101	0.070	NO			
Cartesian Distance to Liquor Store (NEAR_DIST)	1.833	1.822	0.000	YES	0.000	1.000	0.659
Travel Distance to Liquor Store (NETWORK_DIST)	1.724	1.713	0.000	YES	0.000	1.000	0.000
Number of Nearby Cases (NEARBY_CASE)	1.387	1.258	0.000	YES	0.022	1.022	0.143
Corporate Ownership (CORP_OWNED)	1.105	1.105	0.752	NO			
Constant					1.602	4.961	0.000

Appendix D Example GWR4 Output

```
*******************************************************************************
*             Semiparametric Geographically Weighted Regression               *
*                        Release 1.0.80 (GWR 4.0.80)                          *
*                              12 March 2014                                  *
*                  (Originally coded by T. Nakaya: 1 Nov 2009)                *
*                                                                             *
*              Tomoki Nakaya(1), Martin Charlton(2), Paul Lewis(2),           *
*         Jing Yao (3), A. Stewart Fotheringham (3), Chris Brunsdon (2)       *
*                          (c) GWR4 development team                          *
* (1) Ritsumeikan University, (2) National University of Ireland, Maynooth,  *
*         (3) University of St. Andrews                                       *
*******************************************************************************

Program began at 11/27/2015 10:44:01 AM

*******************************************************************************
Session: AREA3_GWLR
Session control file: C:\Users\Matthew\Documents\THESIS PROJECT 2\FULL_DATA_OUT
*******************************************************************************
Data filename: C:\Users\Matthew\Documents\THESIS PROJECT 2\FULL_DATA_OUTPUT\GWR
Number of areas/points: 2198

Model settings--------------------------------
Model type: Logistic
Geographic kernel: adaptive bi-square
Method for optimal bandwidth search: Golden section search
Criterion for optimal bandwidth: AICc
Number of varying coefficients: 8
Number of fixed coefficients:   0

Modelling options-----------------------------------
Standardisation of independent variables: OFF
Testing geographical variability of local coefficients: OFF
Local to Global Variable selection: OFF
Global to Local Variable selection: OFF
Prediction at non-regression points: OFF

Variable settings-----------------------------------
Areal key is not specified
Easting (x-coord): field4 : LONGITUDE
Northing (y-coord): field3: LATITUDE
Lat-lon coordinates: Spherical distance
Dependent variable: field32: PROACTIVE_DUM
Offset variable is not specified
Intercept: varying (Local) intercept
Independent variable with varying (Local) coefficient: field10: FLOOR_AREA
Independent variable with varying (Local) coefficient: field12: NOSTORY_DUM
Independent variable with varying (Local) coefficient: field20: DAYS_TO_VI
Independent variable with varying (Local) coefficient: field21: CASE_COUNT
Independent variable with varying (Local) coefficient: field26: NEAR_DIST
Independent variable with varying (Local) coefficient: field27: NETWORK_DI
Independent variable with varying (Local) coefficient: field28: NEARBY_CAS
*******************************************************************************
```

```
*************************************************************************
  Global regression result
*************************************************************************
  < Diagnostic information >
Number of parameters:                    8
Deviance:                         2415.060237
Classic AIC:                      2431.060237
AICc:                             2431.126020
BIC/MDL:                          2476.622662
Percent deviance explained           0.173466

Variable                  Estimate    Standard Error      z(Est/SE)        Exp(Est)
------------------- -------------- -------------- -------------- --------------
Intercept                 2.038339       0.348676        5.845934        7.677845
FLOOR_AREA               -0.623133       0.612522       -1.017324        0.536262
NOSTORY_DUM              -0.436814       0.124091       -3.520115        0.646091
DAYS_TO_VI               -0.000487       0.000035      -14.077194        0.999513
CASE_COUNT               -0.087511       0.068742       -1.273042        0.916208
NEAR_DIST                -0.000018       0.000041       -0.441124        0.999982
NETWORK_DI               -0.000123       0.000030       -4.073953        0.999877
NEARBY_CAS                0.022249       0.015178        1.465869        1.022498

*************************************************************************
  GWR (Geographically weighted regression) bandwidth selection
*************************************************************************

Bandwidth search <golden section search>
  Limits: 66,  2198
 Golden section search begins...
 Initial values
  pL            Bandwidth:   220.727 Criterion:   2363.251
  p1            Bandwidth:   975.978 Criterion:   2338.606
  p2            Bandwidth:  1442.749 Criterion:   2350.168
  pU            Bandwidth:  2198.000 Criterion:   2349.536
 iter    1 (p1) Bandwidth:   975.978 Criterion:   2338.606 Diff:    466.771
 iter    2 (p2) Bandwidth:   975.978 Criterion:   2338.606 Diff:    288.480
 iter    3 (p1) Bandwidth:   975.978 Criterion:   2338.606 Diff:    178.291
 iter    4 (p1) Bandwidth:   865.788 Criterion:   2338.200 Diff:    110.190
 iter    5 (p2) Bandwidth:   865.788 Criterion:   2338.200 Diff:     68.101
 iter    6 (p1) Bandwidth:   865.788 Criterion:   2338.200 Diff:     42.089
 iter    7 (p2) Bandwidth:   865.788 Criterion:   2338.200 Diff:     26.012
 iter    8 (p2) Bandwidth:   881.865 Criterion:   2338.021 Diff:     16.076
 iter    9 (p2) Bandwidth:   891.801 Criterion:   2337.982 Diff:      9.936
 iter   10 (p2) Bandwidth:   897.941 Criterion:   2337.894 Diff:      6.141
 iter   11 (p1) Bandwidth:   897.941 Criterion:   2337.894 Diff:      3.795
 iter   12 (p2) Bandwidth:   897.941 Criterion:   2337.894 Diff:      2.346
 iter   13 (p1) Bandwidth:   897.941 Criterion:   2337.894 Diff:      1.450
Best bandwidth size 897.000
Minimum AICc     2337.894

*************************************************************************
```

```
*********************************************************************************
  GWR (Geographically weighted regression) result
*********************************************************************************
  Bandwidth and geographic ranges
Bandwidth size:                  897.941202
Coordinate                Min              Max           Range
--------------- -------------- -------------- ---------------
X-coord             -117.412538    -117.277648       12.358629
Y-coord               34.477714      34.556333        8.742099
 (Note: Ranges are shown in km.)

  Diagnostic information
Effective number of parameters (model: trace(S)):                          39.260471
Effective number of parameters (variance: trace(S'WSW^-1)):                30.806895
Degree of freedom (model: n - trace(S)):                                 2158.739529
Degree of freedom (residual: n - 2trace(S) + trace(S'WSW^-1)):           2150.285954
Deviance:                             2257.908274
Classic AIC:                          2336.429215
AICc:                                 2337.894308
BIC/MDL:                              2560.029496
Percent deviance explained               0.227250

************************************************************
 << Geographically varying (Local) coefficients >>
************************************************************
Estimates of varying coefficients have been saved in the following file.
    Listwise output file: C:\Users\Matthew\Documents\THESIS PROJECT 2\FULL_DATA_OUTPU

Summary statistics for varying (Local) coefficients
Variable                      Mean             STD
------------------- -------------- ---------------
Intercept                 2.136204        0.878794
FLOOR_AREA               -1.376446        2.800576
NOSTORY_DUM              -0.048784        0.227288
DAYS_TO_VI               -0.000508        0.000115
CASE_COUNT               -0.109810        0.197216
NEAR_DIST                -0.000020        0.000111
NETWORK_DI               -0.000131        0.000133
NEARBY_CAS                0.051318        0.042055

Variable                         Min              Max           Range
------------------- -------------- ---------------
Intercept                -0.384426        3.291253        3.675679
FLOOR_AREA               -5.184424        3.400054        8.584479
NOSTORY_DUM              -0.488761        0.274475        0.763236
DAYS_TO_VI               -0.000725       -0.000367        0.000357
CASE_COUNT               -0.428870        0.115705        0.544575
NEAR_DIST                -0.000264        0.000231        0.000495
NETWORK_DI               -0.000331        0.000077        0.000409
NEARBY_CAS               -0.026916        0.142533        0.169449
```

```
Variable                Lwr Quartile          Median    Upr Quartile
-------------------- -------------- -------------- --------------
Intercept                   1.567271        2.097494        3.061563
FLOOR_AREA                 -5.038949       -1.712994        1.149981
NOSTORY_DUM                -0.224727       -0.029779        0.267644
DAYS_TO_VI                 -0.000573       -0.000497       -0.000376
CASE_COUNT                 -0.301174       -0.093965        0.112200
NEAR_DIST                  -0.000100        0.000015        0.000049
NETWORK_DI                 -0.000328       -0.000101       -0.000064
NEARBY_CAS                  0.007795        0.061018        0.076409

Variable              Interquartile R     Robust STD
-------------------- --------------- ---------------
Intercept                   1.494292        1.107704
FLOOR_AREA                  6.188930        4.587791
NOSTORY_DUM                 0.492371        0.364989
DAYS_TO_VI                  0.000197        0.000146
CASE_COUNT                  0.413374        0.306430
NEAR_DIST                   0.000149        0.000110
NETWORK_DI                  0.000264        0.000196
NEARBY_CAS                  0.068615        0.050863
 (Note: Robust STD is given by (interquartile range / 1.349) )

********************************************************************************
 GWR Analysis of Deviance Table
********************************************************************************
Source                 Deviance        DOF     Deviance/DOF
------------ ------------------ ---------- ----------------
Global model           2415.060   2190.000            1.103
GWR model              2257.908   2150.286            1.050
Difference              157.152     39.714            3.957

********************************************************************************
Program terminated at 11/27/2015 10:46:12 AM
```

	Multicollinearity		Predictors		Coefficients & Relationships		
Variable Name and (Alias)	Initial VIF Value	Final VIF Value	Predictor Significance	Keep In Model?	Variable Coefficient	Vaiable Odds Ratio - Exp(B)	Odds Ratio Significance
AREA 1							
Lot Size (LOTSQFT)	4.116	1.520	0.022	YES	0.000	1.000	0.335
Floor Area (FLOOR_AREA)	1.889	1.718	0.000	YES	-3.978	0.019	0.025
Number of Stories (NOSTORY)	2.842	1.475	0.000	YES	-0.140	0.869	0.667
Length of Ownership (LENGTH_OWN)	1.184	1.088	0.009	YES	0.000	0.039	1.000
Structure Age (STRUCT_AGE)	1.047	1.020	0.208	NO			

Total Assessed Value (TOTAL_VALU)	7.646	REMOVED					
Value per Sq Ft (VALUE_PSF)	4.186	1.215	0.698	YES	0.001	1.001	0.928
Code Case Yes/No (CODECASE)	5.955	REMOVED					
Days to Previous Violation (DAYS_TO_VI)	7.812	2.314	0.000	YES	0.000	1.000	0.000
Number of Previous Cases (CASE_COUNT)	2.553	2.443	0.000	YES	0.147	0.200	1.158
Tax Defaulted Yes/No (TAX_STATUS)	1.033	1.019	0.208	NO			
Occupancy (OCCUPANCY)	1.152	1.035	0.934	NO			
Cartesian Distance to Liquor Store (NEAR_DIST)	5.772	REMOVED					
Travel Distance to Liquor Store (NETWORK_DIST)	1.171	1.106	0.000	YES	0.000	1.000	0.000
Number of Nearby Cases (NEARBY_CASE)	3.229	1.966	0.000	YES	0.059	1.061	0.069
Corporate Ownership (CORP_OWNED)	1.172	1.154	0.504	NO			
Constant					2.396	10.976	0.054
AREA 2							
Lot Size (LOTSQFT)	4.116	REMOVED					
Floor Area (FLOOR_AREA)	1.889	1.718	0.265	NO			
Number of Stories (NOSTORY)	2.842	1.475	0.820	NO			
Length of Ownership (LENGTH_OWN)	1.184	1.088	0.811	NO			
Structure Age (STRUCT_AGE)	1.047	1.020	0.640	NO			
Total Assessed Value (TOTAL_VALU)	7.646	REMOVED					
Value per Sq Ft (VALUE_PSF)	4.186	1.215	0.048	YES	0.001	1.001	0.928
Code Case Yes/No (CODECASE)	5.955	REMOVED					
Days to Previous Violation (DAYS_TO_VI)	7.812	2.314	0.000	YES	-0.001	0.999	0.000
Number of Previous Cases (CASE_COUNT)	2.553	2.443	0.000	YES	0.067	1.069	0.695
Tax Defaulted Yes/No (TAX_STATUS)	1.033	1.019	0.643	NO			

Occupancy (OCCUPANCY)	1.152	1.035	0.206	NO			
Cartesian Distance to Liquor Store (NEAR_DIST)	5.772	REMOVED					
Travel Distance to Liquor Store (NETWORK_DIST)	1.171	1.106	0.845	NO			
Number of Nearby Cases (NEARBY_CASE)	3.229	1.966	0.000	YES	0.001	1.001	0.973
Corporate Ownership (CORP_OWNED)	1.172	1.154	0.616	NO			
Constant					1.261	3.527	0.154
AREA 3							
Lot Size (LOTSQFT)	5.881	REMOVED					
Floor Area (FLOOR_AREA)	8.405	1.726	0.247	NO			
Number of Stories (NOSTORY)	2.656	1.752	0.156	NO			
Length of Ownership (LENGTH_OWN)	1.202	1.170	0.026	YES	0.000	1.000	0.109
Structure Age (STRUCT_AGE)	1.006	1.004	0.775	NO			
Total Assessed Value (TOTAL_VALU)	7.826	REMOVED					
Value per Sq Ft (VALUE_PSF)	6.509	1.345	0.452	NO			
Code Case Yes/No (CODECASE)	9.396	REMOVED					
Days to Previous Violation (DAYS_TO_VI)	2.843	2.843	0.000	YES	-0.001	0.999	0.000
Number of Previous Cases (CASE_COUNT)	2.823	1.115	0.000	YES	-0.250	0.779	0.033
Tax Defaulted Yes/No (TAX_STATUS)	1.005	1.004	0.357	NO			
Occupancy (OCCUPANCY)	1.161	1.149	0.003	YES	0.071	1.074	0.668
Cartesian Distance to Liquor Store (NEAR_DIST)	1.191	1.191	0.015	YES	0.000	1.000	0.201
Travel Distance to Liquor Store (NETWORK_DIST)	1.184	1.181	0.003	YES	0.000	1.000	0.015
Number of Nearby Cases (NEARBY_CASE)	1.120	1.082	0.000	YES	0.034	1.034	0.140
Corporate Ownership (CORP_OWNED)	1.110	1.114	0.278	NO			

Constant					2.114	8.284	0.000
RANDOM SAMPLE 1							
Lot Size (LOTSQFT)	6.066	REMOVED					
Floor Area (FLOOR_AREA)	8.364	1.541	0.434	NO			
Number of Stories (NOSTORY)	2.178	1.552	0.071	NO			
Length of Ownership (LENGTH_OWN)	1.117	1.111	0.314	NO			
Structure Age (STRUCT_AGE)	1.033	1.031	0.172	NO			
Total Assessed Value (TOTAL_VALU)	7.542	REMOVED					
Value per Sq Ft (VALUE_PSF)	5.998	1.262	0.351	NO			
Code Case Yes/No (CODECASE)	9.091	REMOVED					
Days to Previous Violation (DAYS_TO_VI)	2.736	2.732	0.000	YES	0.000	1.000	0.000
Number of Previous Cases (CASE_COUNT)	2.820	2.803	0.000	YES	-0.016	0.984	0.919
Tax Defaulted Yes/No (TAX_STATUS)	1.034	1.033	0.927	NO			
Occupancy (OCCUPANCY)	1.140	1.130	0.549	NO			
Cartesian Distance to Liquor Store (NEAR_DIST)	1.746	1.738	0.063	NO			
Travel Distance to Liquor Store (NETWORK_DIST)	1.694	1.694	0.095	NO			
Number of Nearby Cases (NEARBY_CASE)	1.338	1.198	0.000	YES	0.060	1.062	0.035
Corporate Ownership (CORP_OWNED)	1.088	1.088	0.551	NO			
Constant					0.351	1.421	0.369
RANDOM SAMPLE 2							
Lot Size (LOTSQFT)	7.635	REMOVED					
Floor Area (FLOOR_AREA)	10.123	1.530	0.149	NO			
Number of Stories (NOSTORY)	2.308	1.581	0.001	YES	0.839	2.315	0.000
Length of Ownership (LENGTH_OWN)	1.136	1.127	0.135	NO			
Structure Age (STRUCT_AGE)	1.024	1.020	0.121	NO			

Total Assessed Value (TOTAL_VALU)	8.783	REMOVED					
Value per Sq Ft (VALUE_PSF)	6.818	1.291	0.564	NO			
Code Case Yes/No (CODECASE)	9.475	REMOVED					
Days to Previous Violation (DAYS_TO_VI)	10.677	2.253	0.000	YES	-0.001	0.999	0.000
Number of Previous Cases (CASE_COUNT)	2.353	2.344	0.000	YES	-0.168	0.846	0.210
Tax Defaulted Yes/No (TAX_STATUS)	1.054	1.046	0.303	NO			
Occupancy (OCCUPANCY)	1.140	1.071	0.892	NO			
Cartesian Distance to Liquor Store (NEAR_DIST)	1.821	1.804	0.048	YES	0.000	1.000	0.926
Travel Distance to Liquor Store (NETWORK_DIST)	1.659	1.633	0.010	YES	0.000	1.000	0.058
Number of Nearby Cases (NEARBY_CASE)	1.456	1.302	0.000	YES	0.016	1.016	0.637
Corporate Ownership (CORP_OWNED)	1.112	1.112	0.918	NO			
Constant					1.225	3.405	0.034
RANDOM SAMPLE 3							
Lot Size (LOTSQFT)	1.082	1.510	0.336	NO			
Floor Area (FLOOR_AREA)	1.807	2.170	0.001	YES	-2.158	0.116	0.094
Number of Stories (NOSTORY)	2.075	1.676	0.004	YES	0.399	1.491	0.112
Length of Ownership (LENGTH_OWN)	1.172	1.170	0.740	NO			
Structure Age (STRUCT_AGE)	1.044	1.037	0.033	YES	0.007	1.007	0.359
Total Assessed Value (TOTAL_VALU)	3.248	REMOVED					
Value per Sq Ft (VALUE_PSF)	2.805	1.311	0.796	NO			
Code Case Yes/No (CODECASE)	9.376	REMOVED					
Days to Previous Violation (DAYS_TO_VI)	10.516	2.092	0.000	YES	0.000	1.000	0.000
Number of Previous Cases (CASE_COUNT)	2.223	2.196	0.000	YES	-0.091	0.913	0.446
Tax Defaulted Yes/No (TAX_STATUS)	1.053	1.047	0.153	NO			

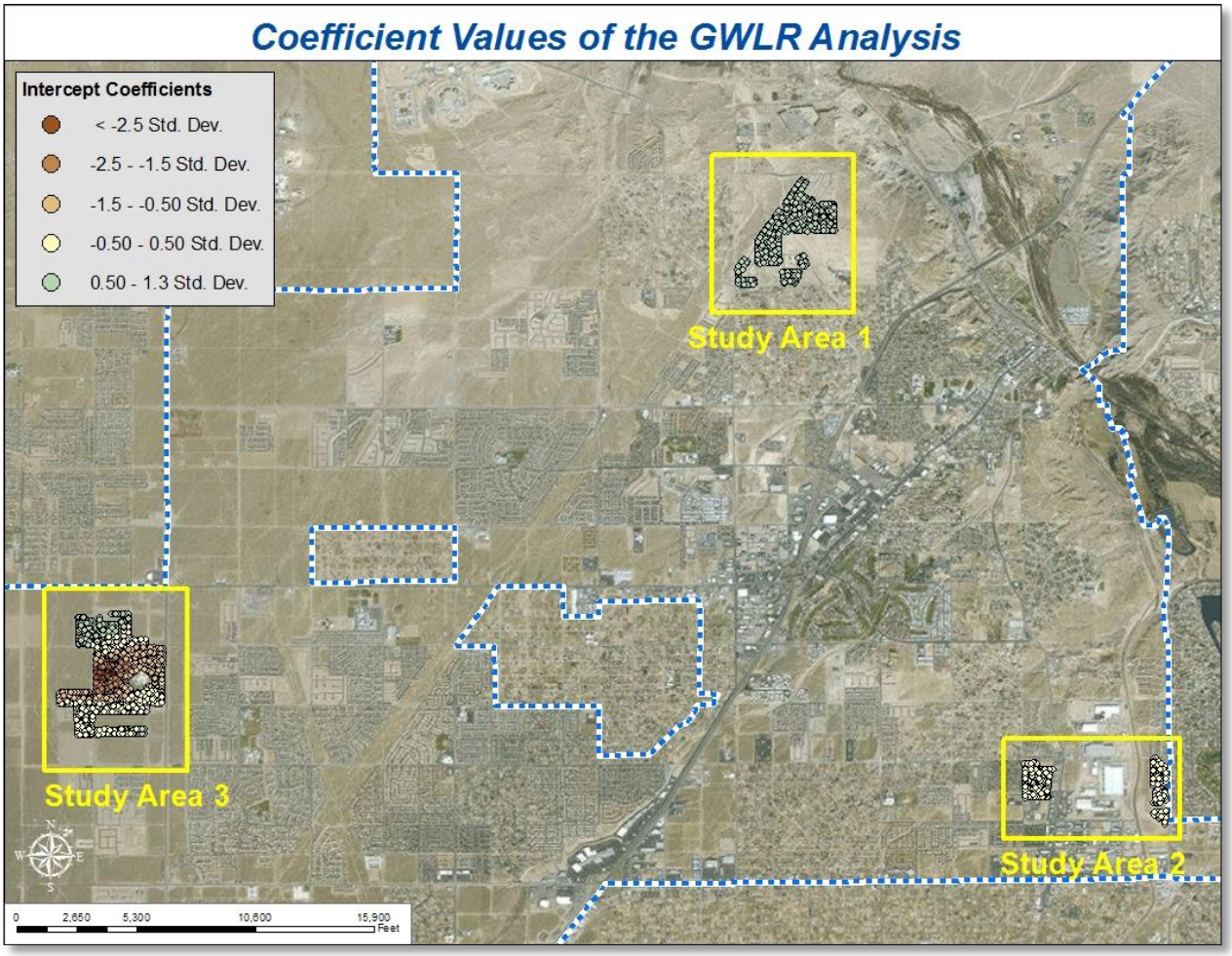
Occupancy (OCCUPANCY)	1.207	1.148	0.314	NO			
Cartesian Distance to Liquor Store (NEAR_DIST)	1.969	1.953	0.022	YES	0.000	1.000	0.685
Travel Distance to Liquor Store (NETWORK_DIST)	1.804	1.776	0.000	YES	0.000	1.000	0.009
Number of Nearby Cases (NEARBY_CASE)	1.386	1.293	0.001	YES	0.019	1.019	0.554
Corporate Ownership (CORP_OWNED)	1.085	1.075	0.136	NO			
Constant					2.119	8.324	0.011
RANDOM SAMPLE 4							
Lot Size (LOTSQFT)	7.351	REMOVED					
Floor Area (FLOOR_AREA)	8.822	1.557	0.163	NO			
Number of Stories (NOSTORY)	2.082	1.573	0.032	YES	0.598	1.818	0.008
Length of Ownership (LENGTH_OWN)	1.226	1.209	0.504	NO			
Structure Age (STRUCT_AGE)	1.018	1.016	0.304	NO			
Total Assessed Value (TOTAL_VALU)	7.917	REMOVED					
Value per Sq Ft (VALUE_PSF)	5.786	1.395	0.025	YES	-0.017	0.983	0.005
Code Case Yes/No (CODECASE)	9.925	REMOVED					
Days to Previous Violation (DAYS_TO_VI)	11.466	2.510	0.000	YES	0.000	1.000	0.000
Number of Previous Cases (CASE_COUNT)	2.698	2.682	0.000	YES	0.012	1.012	0.938
Tax Defaulted Yes/No (TAX_STATUS)	1.052	1.035	0.284	NO			
Occupancy (OCCUPANCY)	1.233	1.164	0.001	YES	-0.168	0.845	0.510
Cartesian Distance to Liquor Store (NEAR_DIST)	1.955	1.952	0.096	NO			
Travel Distance to Liquor Store (NETWORK_DIST)	1.730	1.740	0.006	YES	0.000	1.000	0.013
Number of Nearby Cases (NEARBY_CASE)	1.498	1.329	0.000	YES	0.061	1.063	0.039
Corporate Ownership (CORP_OWNED)	1.103	1.105	0.879	NO			

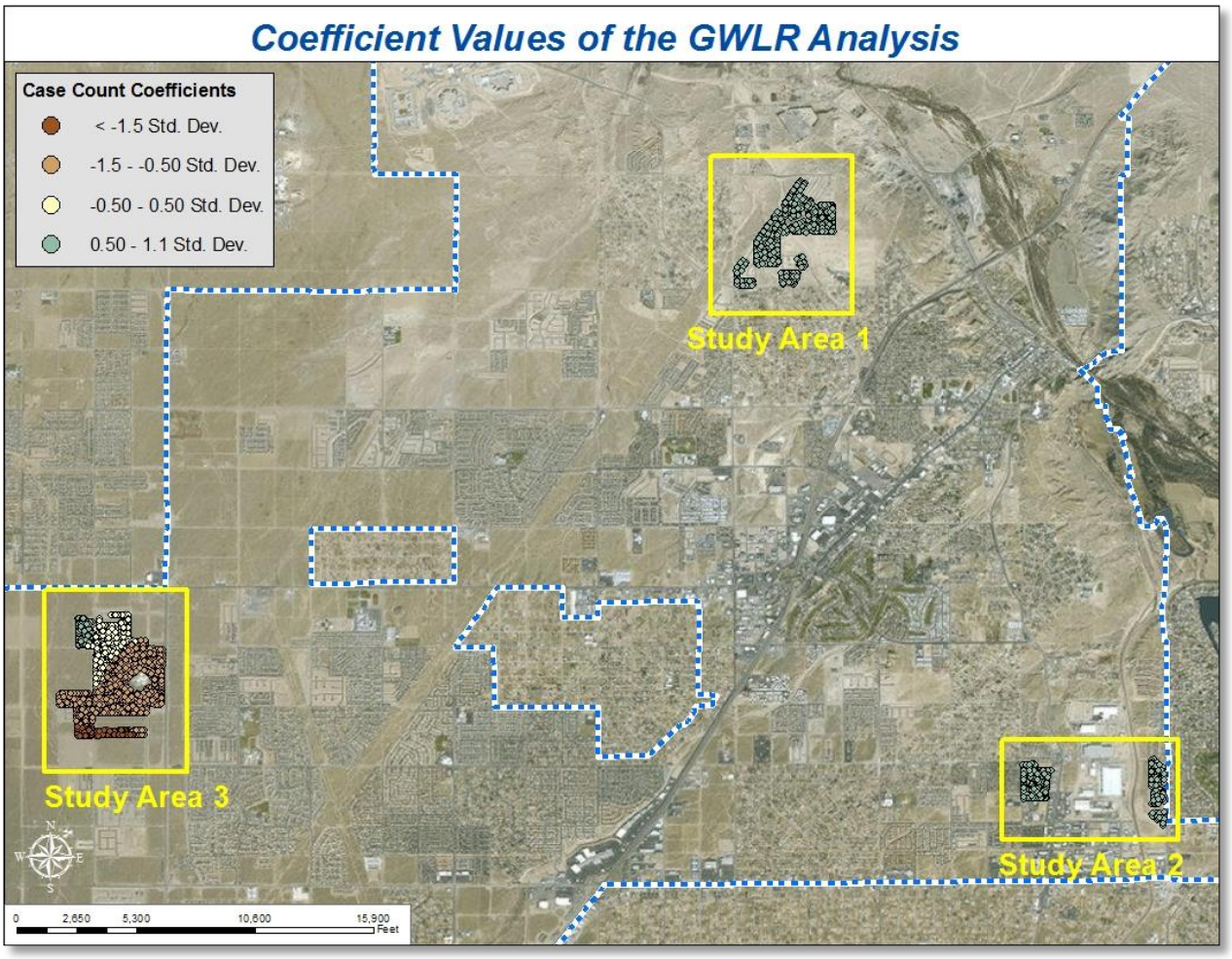
Constant					2.111	8.254	0.004
RANDOM SAMPLE 5							
Lot Size (LOTSQFT)	6.725	REMOVED					
Floor Area (FLOOR_AREA)	9.282	1.547	0.016	YES	-0.479	0.620	0.727
Number of Stories (NOSTORY)	2.291	1.592	0.000	YES	0.602	1.826	0.025
Length of Ownership (LENGTH_OWN)	1.175	1.159	0.698	NO			
Structure Age (STRUCT_AGE)	1.023	1.021	0.173	NO			
Total Assessed Value (TOTAL_VALU)	7.295	REMOVED					
Value per Sq Ft (VALUE_PSF)	5.160	1.250	0.573	NO			
Code Case Yes/No (CODECASE)	7.867	REMOVED					
Days to Previous Violation (DAYS_TO_VI)	9.722	2.598	0.000	YES	-0.001	0.999	0.000
Number of Previous Cases (CASE_COUNT)	2.613	2.595	0.000	YES	-0.158	0.854	0.320
Tax Defaulted Yes/No (TAX_STATUS)	1.020	1.018	0.423	NO			
Occupancy (OCCUPANCY)	1.167	1.110	0.003	YES	-0.170	0.844	0.518
Cartesian Distance to Liquor Store (NEAR_DIST)	1.888	1.873	0.010	YES	0.000	1.000	0.864
Travel Distance to Liquor Store (NETWORK_DIST)	1.802	1.771	0.001	YES	0.000	1.000	0.101
Number of Nearby Cases (NEARBY_CASE)	1.320	1.220	0.000	YES	0.055	1.056	0.086
Corporate Ownership (CORP_OWNED)	1.115	1.113	0.248	NO			
Constant					1.414	4.111	0.110
COMBINED AREAS							
Lot Size (LOTSQFT)	2.682	REMOVED					
Floor Area (FLOOR_AREA)	4.073	1.551	0.000	YES	-0.623	0.536	0.309
Number of Stories (NOSTORY)	2.144	1.580	0.000	YES	0.437	1.548	0.000
Length of Ownership (LENGTH_OWN)	1.157	1.139	0.503	NO			
Structure Age (STRUCT_AGE)	1.002	1.002	0.597	NO			

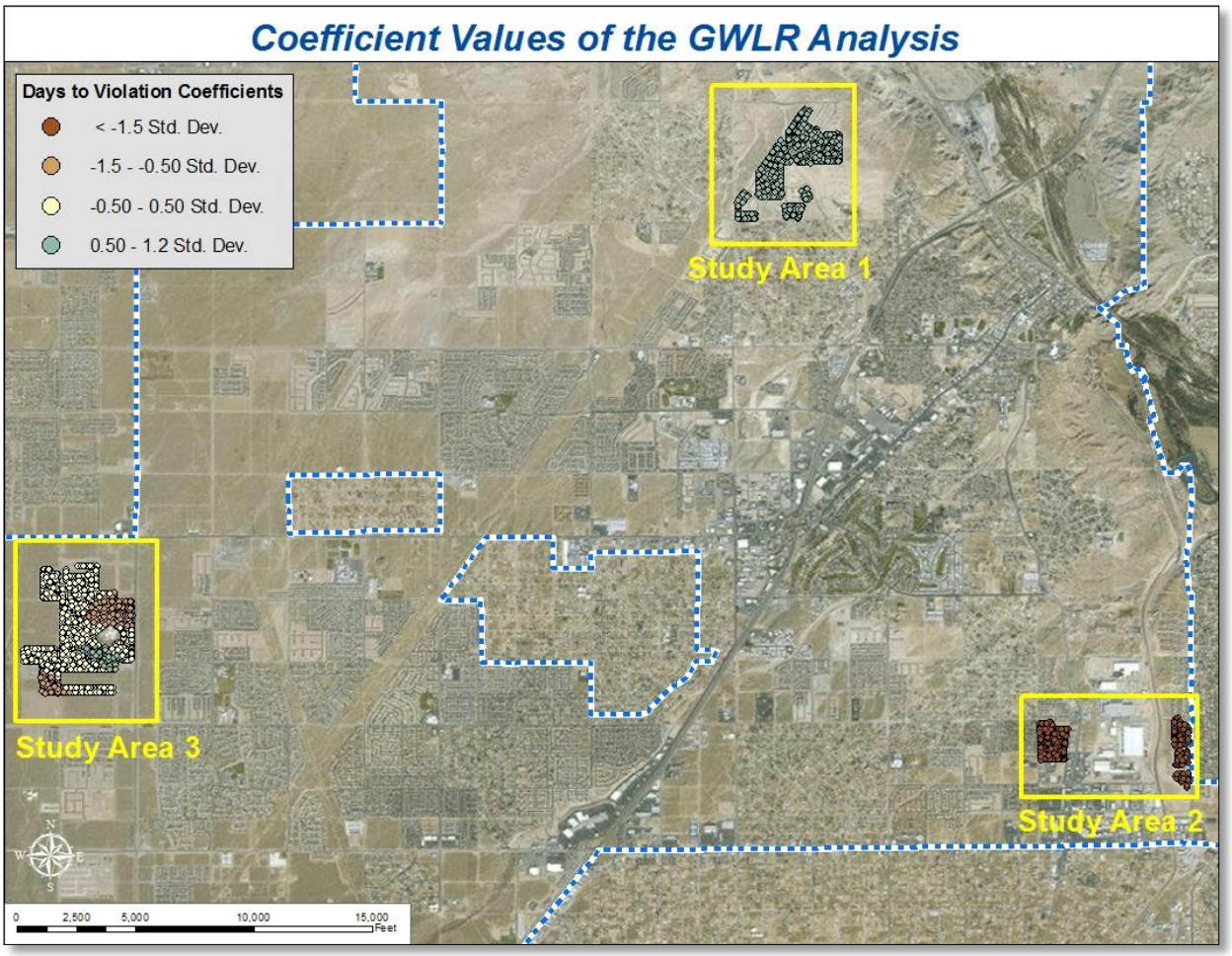
Total Assessed Value (TOTAL_VALU)	4.048	REMOVED					
Value per Sq Ft (VALUE_PSF)	3.286	1.288	0.444	NO			
Code Case Yes/No (CODECASE)	9.106	REMOVED					
Days to Previous Violation (DAYS_TO_VI)	10.761	2.410	0.000	YES	0.000	1.000	0.000
Number of Previous Cases (CASE_COUNT)	2.516	2.495	0.000	YES	-0.088	0.916	0.203
Tax Defaulted Yes/No (TAX_STATUS)	1.010	1.009	0.341	NO			
Occupancy (OCCUPANCY)	1.173	1.101	0.070	NO			
Cartesian Distance to Liquor Store (NEAR_DIST)	1.833	1.822	0.000	YES	0.000	1.000	0.659
Travel Distance to Liquor Store (NETWORK_DIST)	1.724	1.713	0.000	YES	0.000	1.000	0.000
Number of Nearby Cases (NEARBY_CASE)	1.387	1.258	0.000	YES	0.022	1.022	0.143
Corporate Ownership (CORP_OWNED)	1.105	1.105	0.752	NO			
Constant					1.602	4.961	0.000

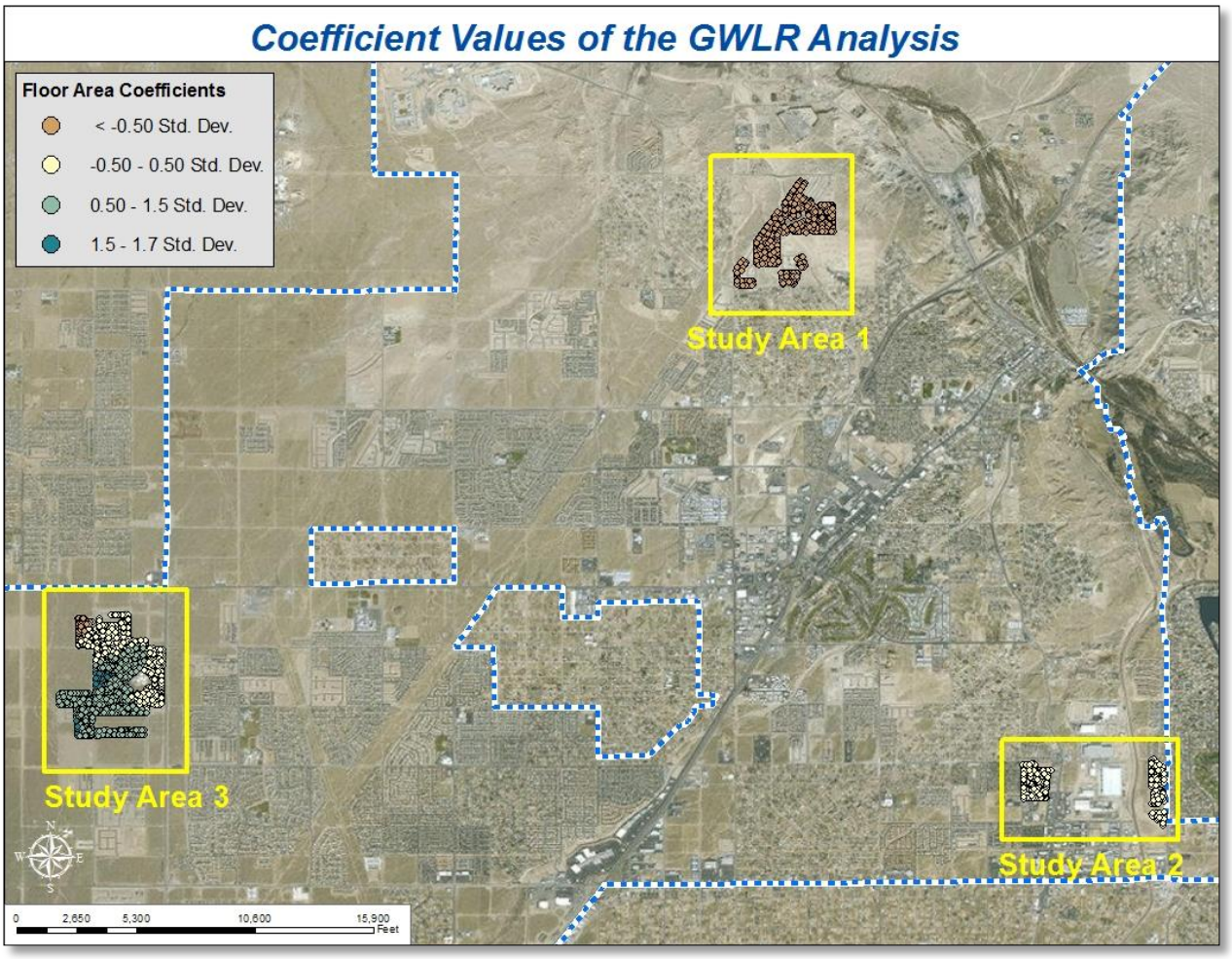
	Y	Z	AA	AB	AC	AD	AE
l	est_NEARBY_CAS	se_NEARBY_CAS	t_NEARBY_CAS	y	yhat	localpdev	Ginfluence
5	0.061437	0.03024	2.0316	0	0.259855	0.233253	0.020395
5	0.061891	0.030443	2.033049	0	0.194053	0.233919	0.012463
5	0.061295	0.03024	2.02699	0	0.165848	0.234563	0.012202
7	0.061786	0.030408	2.031901	0	0.179516	0.233912	0.0144
7	0.061332	0.030134	2.03528	0	0.056753	0.232825	0.006283
1	0.061693	0.030329	2.034135	0	0.156526	0.232533	0.013331
1	0.061081	0.029994	2.036464	0	0.017388	0.233731	0.002191
1	0.060578	0.029819	2.031533	0	0.097737	0.236018	0.009856
1	0.060546	0.029817	2.030597	0	0.106768	0.236052	0.008406
2	0.06112	0.029993	2.03778	0	0.016599	0.233732	0.002174
8	0.060597	0.02982	2.032089	0	0.081244	0.235998	0.009981
2	0.060511	0.029787	2.031446	0	0.077683	0.236432	0.010231
2	0.06045	0.02978	2.029893	0	0.076252	0.236542	0.008094
7	0.06073	0.029861	2.033741	0	0.08198	0.235456	0.006796
1	0.060488	0.029779	2.031216	0	0.053772	0.236541	0.009059
6	0.0605	0.029783	2.031331	0	0.087608	0.236485	0.010562
4	0.060546	0.029802	2.031574	0	0.057245	0.236234	0.010741
4	0.060527	0.029801	2.03103	0	0.063757	0.236252	0.010368
9	0.060515	0.029801	2.030664	0	0.07972	0.236264	0.010584
3	0.061689	0.030377	2.030756	0	0.142665	0.234027	0.010374
6	0.060525	0.029792	2.031576	0	0.124957	0.236371	0.012804
6	0.060493	0.02979	2.030666	0	0.118536	0.236404	0.010312
1	0.060487	0.02979	2.030479	0	0.070662	0.23641	0.009554
2	0.060481	0.029789	2.030298	0	0.060463	0.236416	0.009465
1	0.06119	0.030013	2.038794	0	0.016742	0.233482	0.002426
9	0.06149	0.030273	2.031193	0	0.187958	0.233359	0.017663
4	0.061704	0.030353	2.032929	0	0.138224	0.232995	0.009538
7	0.060557	0.029802	2.031936	0	0.172728	0.236231	0.028799
1	0.060475	0.029789	2.030117	0	0.19467	0.236421	0.027527
7	0.061125	0.030029	2.035169	0	0.280524	0.232461	0.008811

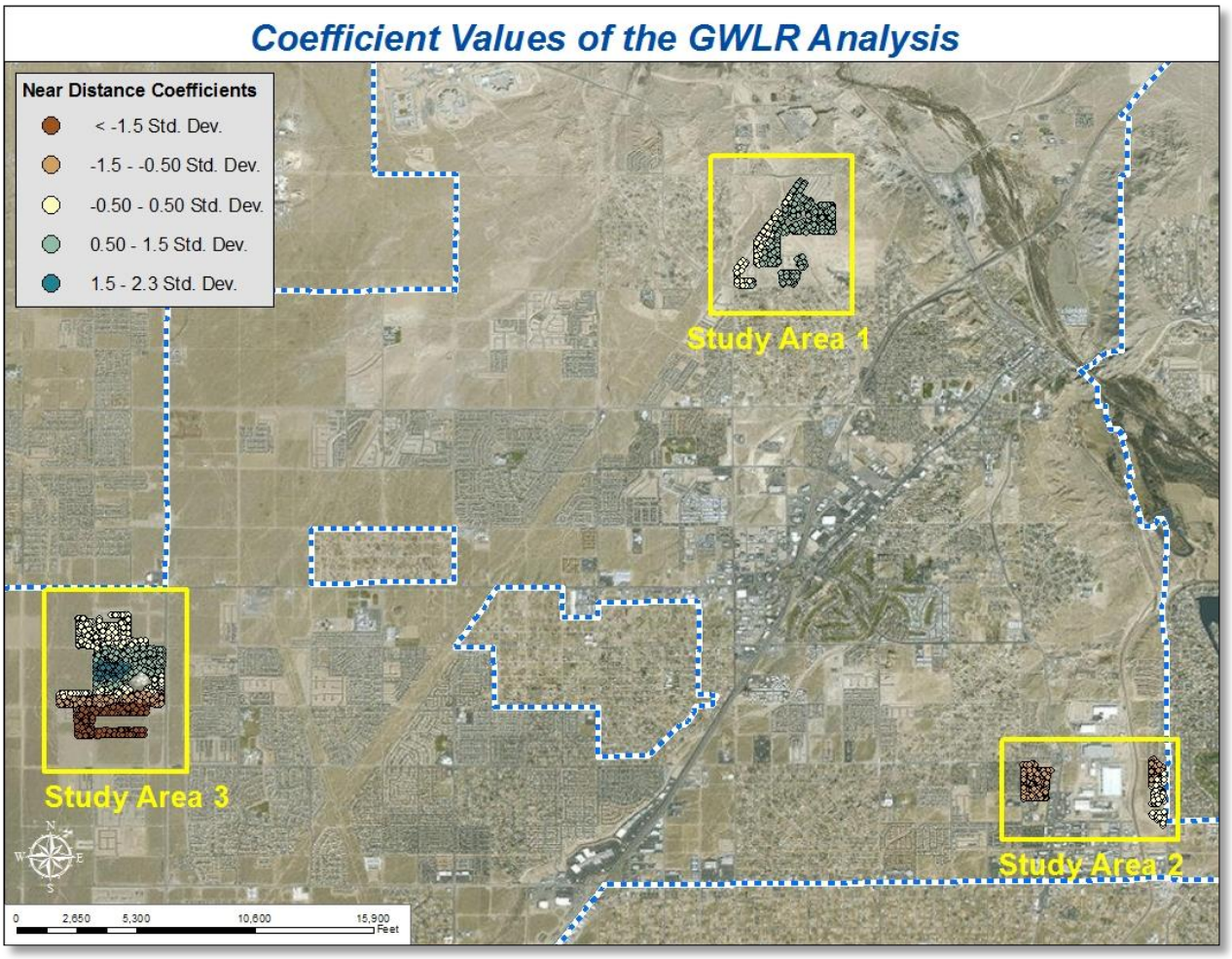
Appendix E GWR4 Coefficient Maps

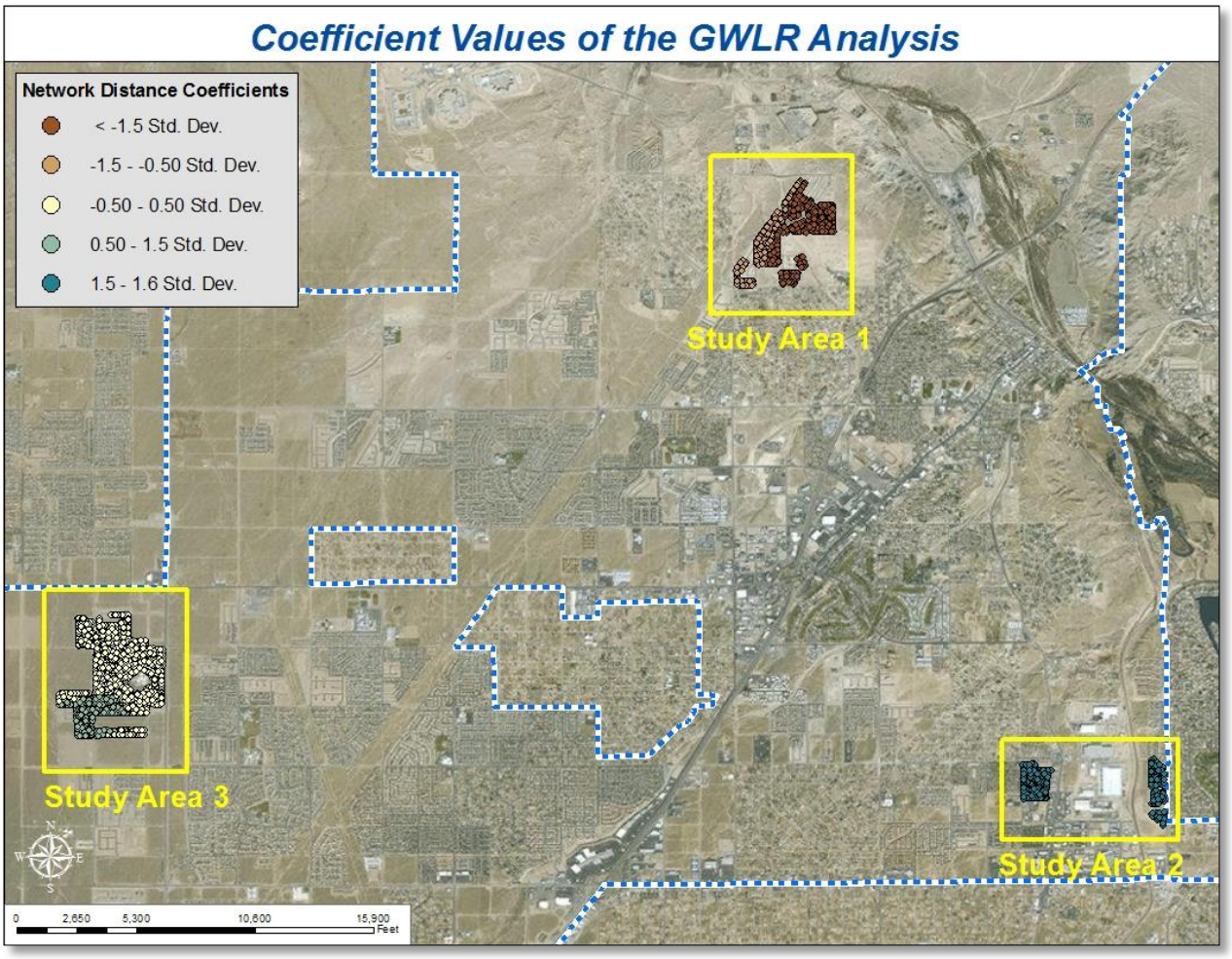


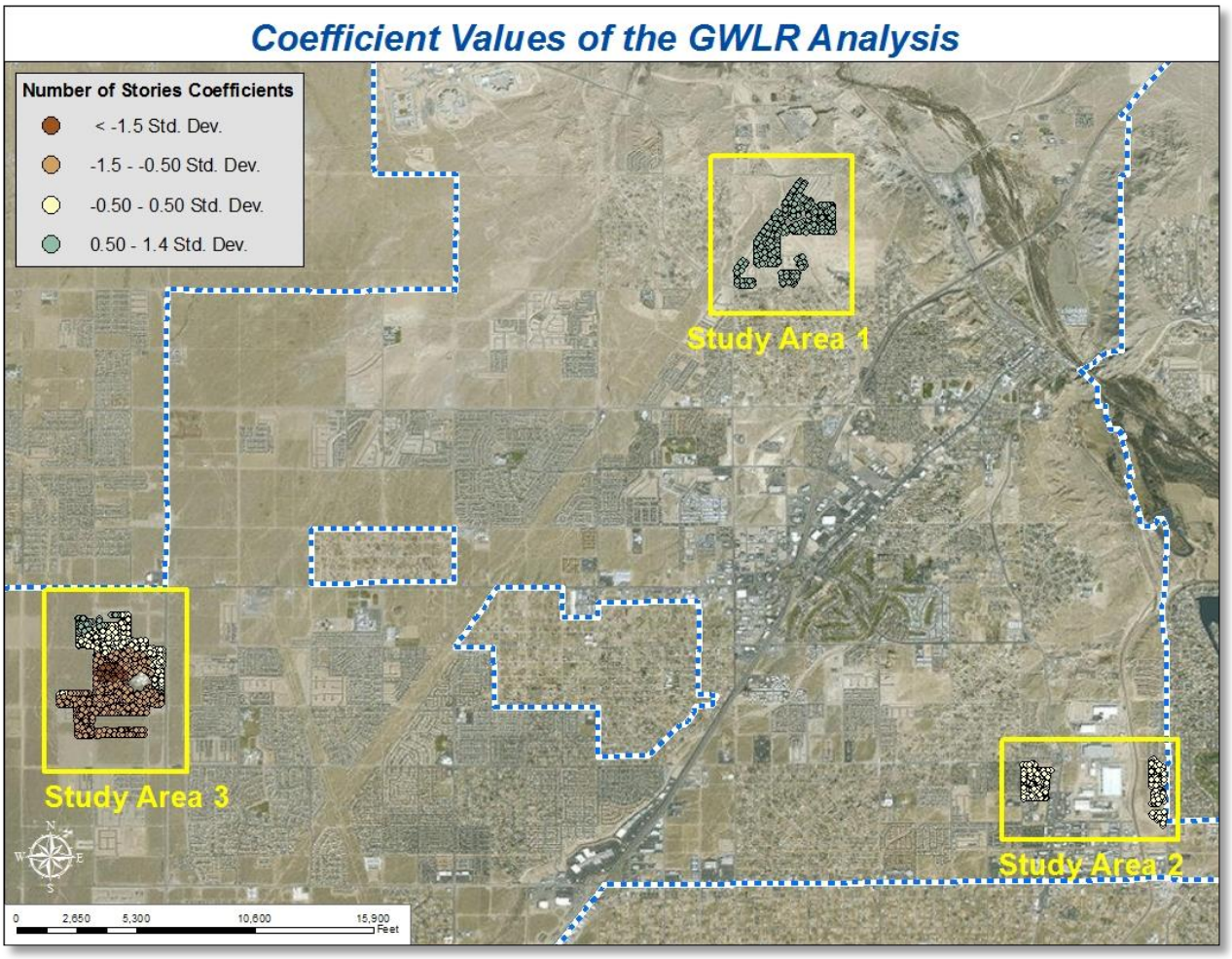

Coefficient Values of the GWLR Analysis
Case Count Coefficients
< -1.5 Std. Dev.
-1.5 - -0.50 Std. Dev.
-0.50 - 0.50 Std. Dev.
0.50 - 1.1 Std. Dev.
Study Area 1
Study Area 3
Study Area 2
0
2,650
5,300
10,600
15,900
Feet

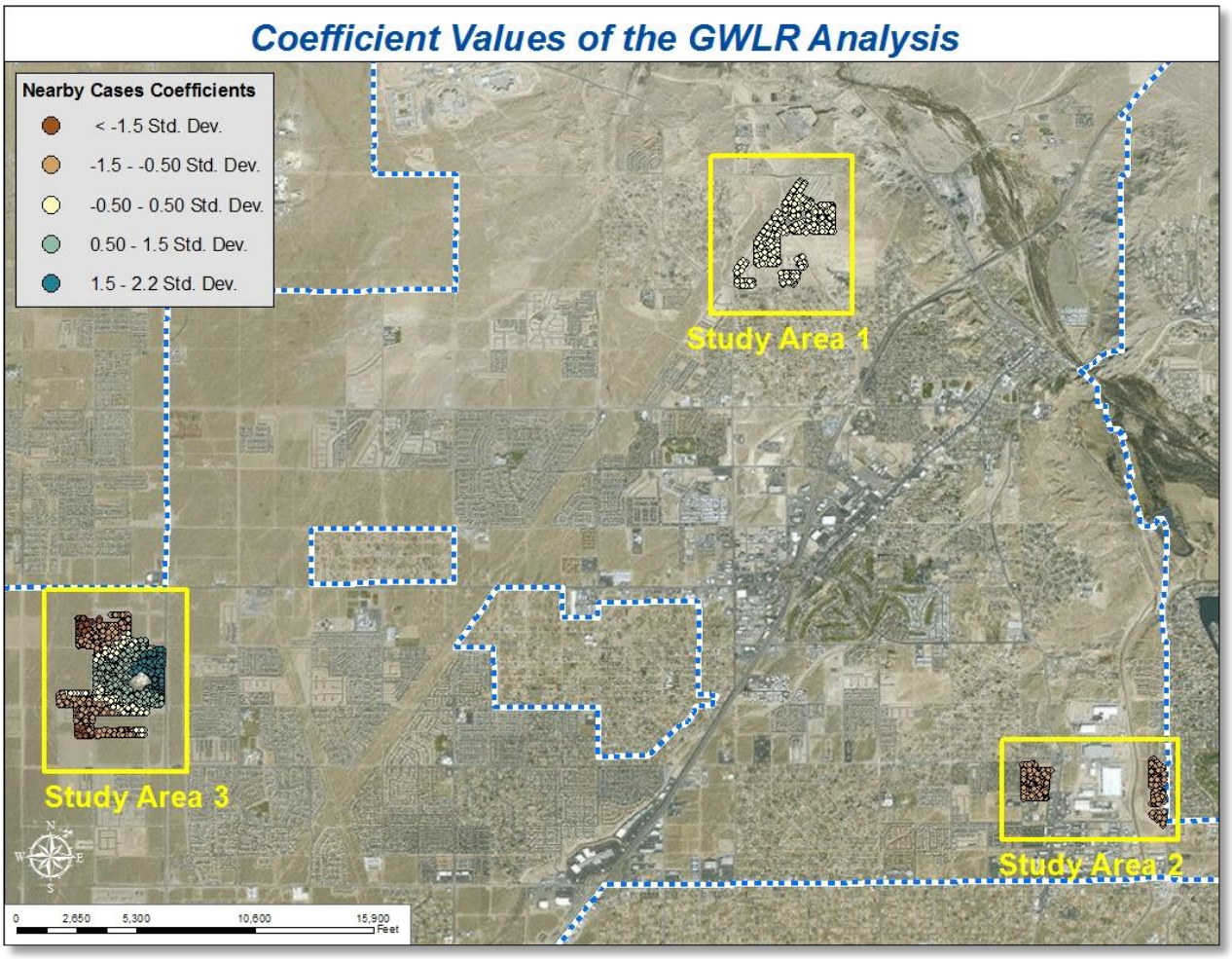

Coefficient Values of the GWLR Analysis
Days to Violation Coefficients
< -1.5 Std. Dev.
-1.5 - -0.50 Std. Dev.
-0.50 - 0.50 Std. Dev.
0.50 - 1.2 Std. Dev.
Study Area 1
Study Area 3
Study Area 2
N
W
E
S
0
2,500
5,000
10,000
15,000
Feet


Coefficient Values of the GWLR Analysis
Floor Area Coefficients
< -0.50 Std. Dev.
-0.50 - 0.50 Std. Dev.
0.50 - 1.5 Std. Dev.
1.5 - 1.7 Std. Dev.
Study Area 1
Study Area 3
Study Area 2
0
2,650
5,300
10,600
15,900
Feet


Coefficient Values of the GWLR Analysis
Near Distance Coefficients
< -1.5 Std. Dev.
-1.5 - -0.50 Std. Dev.
-0.50 - 0.50 Std. Dev.
0.50 - 1.5 Std. Dev.
1.5 - 2.3 Std. Dev.
Study Area 1
Study Area 3
Study Area 2
0
2,650
5,300
10,600
15,900
Feet


Coefficient Values of the GWLR Analysis
Network Distance Coefficients
< -1.5 Std. Dev.
-1.5 - -0.50 Std. Dev.
-0.50 - 0.50 Std. Dev.
0.50 - 1.5 Std. Dev.
1.5 - 1.6 Std. Dev.
Study Area 1
Study Area 3
Study Area 2
0
2,650
5,300
10,600
15,900
Feet


Coefficient Values of the GWLR Analysis
Number of Stories Coefficients
< -1.5 Std. Dev.
-1.5 - -0.50 Std. Dev.
-0.50 - 0.50 Std. Dev.
0.50 - 1.4 Std. Dev.
Study Area 1
Study Area 3
Study Area 2
0
2,650
5,300
10,600
15,900
Feet


Coefficient Values of the GWLR Analysis
Nearby Cases Coefficients
< -1.5 Std. Dev.
-1.5 - -0.50 Std. Dev.
-0.50 - 0.50 Std. Dev.
0.50 - 1.5 Std. Dev.
1.5 - 2.2 Std. Dev.
Study Area 1
Study Area 3
Study Area 2
0
2,650
5,300
10,600
15,900
Feet