DEVELOPMENT OF HOTZONE IDENTIFICATION MODELS FOR SIMULTANEOUS CRIME AND COLLISION REDUCTION

SEUN DANIEL OLUWAJANA

A DISSERTATION SUBMITTED TO
THE FACULTY OF GRADUATE STUDIES
IN PARTIAL FULFILLMENT OF THE REQUIREMENTS
FOR THE DEGREE OF
DOCTOR OF PHILOSOPHY

GRADUATE PROGRAM IN CIVIL ENGINEERING
YORK UNIVERSITY
TORONTO, ONTARIO

December 2018

ABSTRACT

This research contributes to developing macro-level crime and collision prediction models using a new method designed to handle the problem of spatial dependency and over-dispersion in zonal data.

A geographically weighted Poisson regression (GWPR) model and geographically weighted negative binomial regression (GWNBR) model were used for crime and collision prediction. Five years (2009-2013) of crime, collision, traffic, socio-demographic, road inventory, and land use data for Regina, Saskatchewan, Canada were used. The need for geographically weighted models became clear when Moran's I local indicator test showed statistically significant levels of spatial dependency. A bandwidth is a required input for geographically weighted regression models. This research tested two bandwidths: 1) fixed Gaussian and 2) adaptive bi-square bandwidth and investigated which was better suited to the study's database.

Three crime models were developed: violent, non-violent and total crimes. Three collision models were developed: fatal-injury, property damage only and total collisions. The models were evaluated using seven goodness of fit (GOF) tests: 1) Akaike Information Criterion, 2) Bayesian Information Criteria, 3) Mean Square Error, 4) Mean Square Prediction Error, 5) Mean Prediction Bias, and 6) Mean Absolute Deviation. As the seven GOF tests did not produce consistent results, the cumulative residual (CURE) plot was explored.

The CURE plots showed that the GWPR and GWNBR model using fixed Gaussian bandwidth was the better approach for predicting zonal level crimes and collisions in Regina. The GWNBR model has the important advantage that can be used with the empirical Bayes technique to further enhance prediction accuracy.

The GWNBR crime and collision prediction models were used to identify crime and collision hotzones for simultaneous crime and collision reduction in Regina. The research used total collision and total crimes to demonstrate the determination of priority zones for focused law enforcement in Regina. Four enforcement priority zones were identified. These zones cover only 1.4% of the City's area, but account for 10.9% of total crimes and 5.8% of total collisions. The research advances knowledge by examining hotzones at a macro-level and suggesting zones where enforcement and planning for enforcement is likely to be most effective and efficient.

ACKNOWLEDGEMENTS

Firstly, my sincerely appreciation to God for granting me the strength to complete this research. Also, I would like to thank my family for their support throughout the entire course of my studies.

Secondly, I thank and appreciate my supervisor, Dr. Peter Park, for his confidence and courage in accepting an internationally trained student from Nigeria for the Doctor of Philosophy (Ph.D.) program without any doubt in my competence. His guidance and financial support throughout the duration of my studies cannot be overemphasised.

My special thanks and appreciation go to the members of my graduate advisory committee and also to the City of Regina, RPS, and SGI for funding and providing the data used in this research.

I would also like to thank Emmanuel Takyi, Godfred Yeboah, Blessing Obute, Nadeem Abbas, and Samrat Hussein who are my former colleagues and friends at the University of Saskatchewan for their support and encouragement. Also, my appreciation goes to Gbenga Soladoye, Oyedeji Mibiola, and the Ojos and Wahab families for their encouragement. To Crystal Mingue Wang, Erik Nevland, Tanvir Chowdhury and Ravi Rampure, my newly found friends at York University Toronto, thank you so much for your support. Lastly, I must also say to Sindy Mahal and Gillian Moore, thank you for always providing answers to my questions.

Without you all, this research would have been impossible!

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS AND ABBREVIATIONS

AADT	Average Annual Daily Traffic
AIC	Akaike Information Criteria
AIC_C	Corrected Akaike Information Criteria
BIC	Bayesian Information Criteria
COR	City of Regina
CRP	Continuous Risk Profile
CURE	Cumulative Residual
DDACTS	Data Driven Approach to Crimes and Traffic Safety
EB	Empirical Bayes
FI	Fatal-Injury
GOF	Goodness of Fit
GWNBR	Geographically Weighted Negative Binomial Regression
GWPR	Geographically Weighted Poisson Regression
MPB	Mean Prediction Bias
MSE	Mean Square Error
MSPE	Mean Square Prediction Error
NHTSA	National Highway and Transportation Safety Administration

LIST OF ACRONYMS AND ABBREVIATIONS

NIJ	Institute of Justice
PDO	Property Damage Only
RPS	Regina Police Service
RTM	Regression-to-the-Mean
SGI	Saskatchewan Government Insurance
TAIS	Traffic Accident Information System
TAZ	Traffic Analysis Zone
UGRID	Location Identifier
VKMT	Vehicle Kilometers Travelled

LIST OF APPENDICES

CHAPTER 1:INTRODUCTION

1.1 Problem Statement

Keighley (2017) found that there were about 1.9 million police reported crime incidences in 2016 across Canada. This number was approximately 27,700 more than the 2015 crime incidences. Interestingly, the Crime Severity Index (CSI), which is defined as the number of incidences per 100,000 population, for violent crimes across the entire country was unchanged at 75.3 for both 2015 and 2016. Although the total number of police-reported violent crimes decreased slightly (1% reduction) in 2016, certain types of violent incidences increased. These include sexual harassment (combined total of +36% for sexual violation against children and aggravated sexual assault), assault involved with the use of weapons causing bodily harm (+1%), assault (+3%), and aggravated assault (+1%). These statistics demonstrate the need to increase our efforts to reduce the number of violent crimes in our society.

Importantly, the number of reported thefts of $5,000 or under and the number of fraud crimes were 14% higher than in 2015 and contributed to the increase seen in total non-violent crimes while all other types of property crimes remained similar. Overall, the national average CSI for Canada rose from 70.1 in 2015 to 71 in 2016, an increase of about 1%.

Among Canada's provinces, Saskatchewan had a large increase in crime: the CSI increased by 9% for total crimes. Saskatchewan was closely followed by Manitoba (CSI increase of 8%). Among major cities in Canada, Regina, the capital of Saskatchewan, recorded the highest CSI in 2016.

The large number of crimes has impacted our society greatly and is costly for individuals, community and the country as a whole (McCollister et al., 2010; Soh, 2012). In 2014, Canada spent $85 billion on crime control with more than half (about $48 million) going to victim loss to crimes (Easton et al., 2014). While a 1 % increase in the Canada's overall CSI in 2016 (Keighley, 2017) may imply the cost of crimes must also be on the rise, the cost may vary by municipality (Moon et al., 2014). Efforts to minimize the societal cost of crimes need to be on going (Moon et al., 2014).

Other serious societal problems faced by our communities include injuries, loss of life, and loss of property due to traffic collisions. Traffic collisions increase with higher dependency on automobiles and with increases in household vehicle ownership (Kamruzzaman et al., 2013). The easy of use of automobiles coupled with automobiles' advantage of providing easier accessibility to destinations have made vehicles an important part of our lives. This phenomenon has substantially influenced individuals' lifestyles and our urban spaces due to congestion, pollution and, most importantly, increased numbers of traffic collisions (Kim and Ulfarsson, 2008). Fatalities from traffic collisions rank high among the leading causes of preventable death in many countries worldwide (Bhagyaiah and Shrinagesh, 2014). Traffic collisions are among the top causes of accidental death in Canada (Statistics Canada, 2018) and thus can be considered a serious public health concern (Mayhew et al., 2004)

The World Health Organization (WHO, 2018) reported that over 1.25 million people lose their lives every year due to traffic collisions. In particular, traffic-related fatalities are the leading cause of death among young adults. It is estimated that persons aged 15 to 44 account for 48% of deaths involving traffic collisions globally. This estimate is very close to recently released statistics for 2016 (Transport Canada, 2016). Out of the 1,898 reported fatalities involving traffic accidents, persons aged 15 to 44 accounted for about 46 % of total fatalities. Interestingly, the same age group accounted for 53% of the 10,333 serious injuries that occurred during the same period.

Saskatchewan has the highest traffic fatalities rate (10.9 per 100,000 population) for all provinces and territories in Canada. In most countries, road traffic collisions account for about 3% of the Gross Domestic Product (WHO, 2018).

The simultaneous occurrence of crimes and collisions is a real concern for law enforcement authorities. Enforcement authorities are trying to minimize the adverse impact of crimes and collisions to our societies, and proactive policing strategies are being sought. A proactive approach to policing requires predictive techniques that can identify crimes and collision locations or areas and then identify hotspots, i.e. locations/areas where high numbers of incident occurrences are expected.

Various predictive policing strategies have been developed. These include community policing (Gill et al., 2014; Reisig, 2010), hotspot policing (Leigh et al., 2017; Ratcliffe, 2004), smart policing (Coldren et al., 2013), and ubiquitous crime prevention system (UCPC) (Moon et al., 2014). These techniques are known as problem-oriented policing among enforcement authorities and are designed to use rational and evidence-based analysis to address specific problems (Eck and Spelman, 2016). Most of the techniques focus independently on areas generating a high number of crimes and areas generating a high number of collisions. These various techniques are predictive and considered more to result in more effective enforcement tactics than the traditional law enforcement tactics (e.g. 911 calls and random patrols) used by police services (Weisburd et al., 2010).

The concept of predictive policing has been advanced by the National Highway and Traffic Safety Administration (NHTSA) since 2008. Predictive policing takes into consideration the idea that crime and collision occurrences are linked by place related factors such as demography, land use, and socioeconomics. The NHTSA approach is known as the Data Driven Approach to Crimes and Traffic Safety (DDACTS).

DDACTS focuses on simultaneous crime and collision reduction to present a combined set of operational enforcement tactics. The approach stems from the need for enforcement authorities to maximize limited resources by optimizing their services and associated operational costs (Wilson and Heinonen, 2012; Carter and Piza, 2017).

This dissertation assisted in the development of DDACTS enforcement tactics by using advanced rigorous scientific techniques to enhance methods used to determine DDACTS target areas for enforcement.

1.2 Research Motivation

Traditional law enforcement strategies rely on 911 calls and random patrols to bring about a reduction in crimes and traffic collisions. As traditional law enforcement tactics are reactive (crimes and collisions must occur at a place before enforcement is deployed), a reactive approach may not be the most effective way to employ limited enforcement resources to prevent crimes and collisions. The recognition of the limitations of traditional enforcement tactics based on 911 calls and random patrols has led to the development of predictive policing strategies which require the prediction of crime and collision prone locations so that enforcement can be deployed before the occurrence of crimes and collisions.

DDACTS is now a popular and commonly used predictive policing strategy in North America. DDACTS is based on the understanding that both crimes and collisions are influenced by place characteristics. Hence, it has been widely described as a place-based methodology for crime and collision reduction (Alpert and Lum, 2014; Burch and Geraci, 2009; Coppola 2014; Ryberg et al., 2014; Stuster et al., 2010; Takyi et al., 2018).

Over the past decades, studies have shown a relationship between the prevalence of crime and the prevalence of traffic collisions. The recognition of this relationship formed the basis for the DDACTS concept. Research findings have established the connection between crime and road safety using behavioural characteristics of individuals, i.e. the behaviour of persons on and off roadways is likely to be the same. Those who drive aggressively on roadways are likely to be anti-social and likely to be involved in criminal activities. Other research that has linked crime to collisions has employed an assumption that a spatial correlation between these two variables exists (Kuo et al., 2011; Adel et al., 2016). Castillo-Manzano et al. (2015), for instance, reported that the spatial relationship results from common socio-economic and demographic factors. Recent studies have revealed that some places attract individuals who are likely to be involved in criminal activities or get involved in traffic violations that could presumably lead to traffic fatalities (Cook 2012; Sweeney, 2009; Takyi et al., 2018).

These studies point out that there may be some rough areas in a city where a high number of crimes and a high number of collisions occur simultaneously. Proper statistical analysis using selected characteristics can throw additional light on such areas. Eventually, we want to be in a position to predict areas likely to experience a high number of both crimes and collisions. Despite the now demonstrated relationship between crime and traffic collisions, no theory exists in environmental criminology or transportation engineering to establish the expected spatial clusters (Carter and Piza, 2017).

The recognition of the spatial clustering of location of crimes and collisions in our modern society has, however, led to the establishment of DDACTS tactics. DDACTS conducts analysis to determine hotspots and commonly uses kernel density estimates (KDE) to produce a map to visualize the spatial distribution of crimes and collisions across a study area (Gerber, 2014). KDE uses a bandwidth (i.e., a spatial boundary) as an input to determine the area of the planar surface i.e., the 2-dimensional geographic surface considered in evaluating the prevalence of crime and collisions. The KDE technique considers an area or location to be risky if the density of target events (in our case, the number of crimes or collisions) exceeds certain threshold value depending on the optimal bandwidth selected (Yu et al., 2014). Based on the risk potential of the events, hotspots are determined and used to visualize clusters of events over the continuous geographic surface (Vemulapalli, et al., 2017).

It is generally accepted that crime and traffic collisions are seldomly random events in space (Xie and Yan, 2008; Highway Safety Manual, 2010). Occurrences of these events are influenced by certain place-based characteristics. For example, the socio-demographics and the land use of an area significantly influence the number of crime occurrences. These factors coupled with vehicular traffic volume also impact the number of collisions.

The current DDACTS approach, due to its reliance on density estimation, assumes the homogeneity of the 2-dimensional space, i.e., uniformity of space. This assumption can be regarded as problematic as factors that influence crimes or collisions are not homogeneous over geographic space (Xie and Yan, 2008). The sociodemographic and land use characteristics that give rise to crime events over geographic spaces are not uniformly distributed. Also, the vehicle kilometers travel (exposure to traffic collision) are not homogeneously distributed across geographic units. Spatially, the outcomes form clusters consistent with crime and collision occurrences. Areas with poor socio-economic conditions are expected to have high crime incidences. The same applies to exposure (i.e. population) for crimes but most importantly, the same land use can serve as attractors and generators for crimes and collisions at the area level. The assumption of uniformity neglects the social dynamics that create hotspots.

Closely related to the assumption of uniformity in space is the assumption that future crime and collision occurrences will take place near the past event, but the assumption that future events will occur near past ones is not guaranteed. This is because crimes and collisions can be displaced or move to another area due to the influence of various factors such as the socio-economics, demographics, land use, and enforcement. Spatial displacement (referred to as spatial dependency) can contribute to random variation in the number of crimes and collisions. This issue could lead to the misidentification of hotspots.

Current DDACTS techniques also require a large historical database for hotspots to be determined. This requirement implies that DDACTS analysis cannot be done in places without abundant data and that any conclusion drawn on limited data could be ambiguous. Also, the over dependence of DDACTS methods on a large amount of data suggest that crime and collisions must happen repeatedly at a location before that location is identified as a hotspot. In addition, when these hotspots are determined, there are no statistical methods for determining the validity of the prediction made (Gerber, 2014). Furthermore, the hotspots identified are a local description for a geographic zone which cannot be used to describe another geographic zone.

Importantly, we see that current DDACTS methods do not allow for factors that influence crime and collisions to be used in hotspot determination. The methods used do not take into consideration factors that gives rise to certain incidences over the geographic surface (Kloog et al., 2009**)** and results in the static nature of the hotspots identified and the inability to predict future hotspots. While the displacement of crime and collisions can occur due to similarities in the characteristics of the socio-economics and land use of proximal locations, DDACTS' non-consideration of the spatial dependence structure is currently a limitation. Events like crimes and collisions are the consequence of several interacting characteristics of a place, but DDACTS only considers location coordinates in the determination of hotspots and the approach is not predictive in nature.

These comments present some common but largely ignored limitations of the current DDACTS technique. Crime and collisions occurrences are not independent of social factors such as land use, demographics and socio-economics. As a result, one of the major limitations of the existing DDACTS approach is that it is hard to use it to support policy or long-range planning level decisions. Although current DDACTS tactics are designed largely for use as operational level enforcement tactics, improved DDACTS tactics can possibly be used to assist in policy or the long-range planning level of enforcement decisions if DDACTS tactics can be re-formulated to use the predicted occurrence of crime and collisions to determine the hotspots.

The research discussed in this dissertation contributes to the improvement of the existing DDACTS approach such that DDACTS can be used not only for operational level enforcement tactics, but also for policy and planning level enforcement tactics. This could be achieved by using a scenario-based forecast that compares both the base year and the future year scenario using a predictive model. However, a model developed using the base year information is required.

Improved DDACTS tactics that can be used for long range enforcement planning enforcement can be done at a macro-level (zonal-level) only. This is simply due to the current technical ceiling, i.e., no technical tool that can predict the exact location of individual crimes or collisions is available to us. As a result, this dissertation considers hotspot determination at the macro level using aggregated information such as socio-demographics and land use for each zone to identify crime and collision hotspots. Macro-level analysis allows aggregated variables (such as demographics and land use) that affect crime and collisions to be incorporated and used to understand the reasons for hotspot occurrences in different areas. Macro-level analysis also facilitates prediction of potential future hotspots by considering spatial dependency. However, accounting for spatial dependence requires the use of an advanced spatial statistical technique. Nowadays, a popular spatial model for capturing spatial dependence in data is the geographically weighted regression model.

1.3 Research Goal and Objective

The goal of this research was to develop macro-level prediction models to enhance DDACTS methodology and the determination of hotspots that can then be used for long-range enforcement planning and policy development.

The specific objectives of this research were to:

1. Develop macro-level prediction models for crimes and collisions using (a) Geographically Weighted Poisson Regression (GWPR) and (b) Geographically Weighted Negative Binomial Regression (GWNBR) to overcome spatial dependency in crime and collision data;
2. Evaluate the impact of bandwidth choice between (a) fixed Gaussian and (b) adaptive bi-square on the fitting performance of the GWPR and GWNBR models;
3. Suggest a rigorous way of choosing the best fitting GWPR and GWNBR model; and
4. Determine DDACTS hotspots for crimes and collisions at the macro-level to improve the efficiency of enforcement deployment efforts.

1.4 Study Design

This research focuses on the analysis of crime and collisions hotspots, i.e., locations or areas where a high number of crime and collisions incidences is likely to occur. The research considers the development of predictive models of crimes and collisions at the macro-level not only to use at the operational level of enforcement tactics, but also for use in planning and policy level decisions involved in enforcement deployment. The approach can also be regarded as a long-range safety planning exercise for transportation safety professionals.

Long range safety planning models, otherwise known as macro-level prediction models, enable planners to develop empirical tools for proactive safety study. The goal of the development of such models is to provide safety evaluation tools that proactively evaluate the safety levels of regions such that appropriate countermeasures (in this case enforcement) can be provided. The approach involves the development of models at the macro-level to capture the relationship that exists between a variable of interest (i.e., crimes and collisions in our case) and other aggregated level variables that exist at the macro-level (Hadayeghi et al., 2010). Long range planning models such as macro-level crime and collision prediction models help us to understand important variables that influence the occurrence of the variables of interest (Siddiqui et al., 2012).

The application of macro-level prediction models using macro-level variables helps to explain variation in crime and collisions. Variables include demographic, socio-economic, land use and traffic-related variables. These variables are often required when there is a need to relate the safety of an area to macro-level information. Also, such models are useful when there is a need to regularly evaluate the safety level of a region and provide countermeasures to bring about a reduction in the incidences that make an area unsafe (Xu et al., 2014). The use of macro-level models has been gaining attention in crime and collision studies. Statistical methods have also evolved to overcome challenges such as spatial dependency and heterogeneity that might arise with their use. Models that take these issues into consideration offer the advantage of improved precision by considering neighbouring site influences on the occurrence of crime and collisions. Apart from allowing for incorporation of correlates in prediction, macro-level models can help in the identification of hotspots (Dong et al., 2016).

In this dissertation, areas with a high incidence of crime and collisions are referred to as hotzones. Hotzones are areas with a high need for enforcement prioritization. Using the term hotzone was necessary in this dissertation to ensure consistency since the analyses were done at a macro-level using the traffic analysis zone as the geographic unit of analysis.

1.5 Research Scope

The research described in this thesis develops macro-level prediction models that facilitate incorporation of aggregated level variables into the determination of hotzones. The models used are the geographically weighted Poisson regression (GWPR) model and the geographically weighted negative binomial regression (GWNBR) regression.

The research makes a case for the development of spatial models, i.e. the GWPR and GWNBR regression, due to the violation of the assumption of the independence of observation (tested using Moran I). From a scientific perspective, it is expected that the negative binomial regression will underperform as the required assumption was not met. It has been reported that the GWPR models that takes the form of a Poisson regression outperformed with the earlier stated assumption relaxed and fits better than the negative binomial model (Li et al., 2013). It also appears very unlikely that the negative binomial regression will outperform the GWNBR models. As a result, the negative binomial was not compared with the spatial counterparts. The spatial variant known as the GWNBR models was compared with the alternatively used GWPR. Thus, this research demonstrates the sophistication of the GWNBR over the negative binomial regression in terms of the assumption on which it is formulated.

The scope of this research is limited to extending the concept of DDACTS to macro-level hotzone determination to facilitate both operational enforcement and long-range planning and policy formulation. The models developed are similar to travel demand models that could be used for forecasting the future number of collisions at a zonal level by considering anticipated changes that could affect the input variables. However, forecasting future crimes and collisions requires anticipated/projected changes to be included in the input variables. These changes can be determined by city officials and incorporated into the preferred GWNBR models. The models developed provide tools for examining future crime and collision incidence scenarios. Thus, this research shows the use of GWNBR models developed at a macro-level for DDACTS hotzone identification considering our base year (2013) with findings that could be extended to forecast future scenarios.

1.6 Dissertation Structure

The second chapter conducts an extensive review of the literature on current crime and traffic enforcement strategies. The chapter also discusses an alternative approach to hotspot determination using macro-level prediction models. Different levels of aggregations are identified, and the unit preferred by transportation safety professionals is emphasized. Likely challenges that characterize aggregated data analysis are discussed, and methods for overcoming them are evaluated. Chapter two also discusses various statistical approaches that have been explored in crime and collisions prediction including their application, advantages, limitations, and recent advancements.

The third chapter focuses on the approach proposed for the identifications of hotzones for crime and collisions at the macro-level. The chapter explains the source of data, the crime types and collision severity considered, the crime and collision models' methodology, and how areas with a high incidence of crime and collisions can be determined using the predicted values.

Chapters four, five and six are the main components of this dissertation. The chapters review the questions addressed, the data used, and the methods applied. The chapters present a comparison and evaluation of the methods used. The discussion is followed by the conclusions drawn from the analysis.

In particular, chapters four and five discuss the prediction models developed for crime and collisions.

Chapter four is concerned with the collision database and the collision prediction models developed for the City of Regina, Saskatchewan using socio-demographic and land use data aggregated into traffic analysis zones. GWPR and GWNBR models are developed, and the models' predictive performance using two bandwidth types (fixed Gaussian and adaptive bi-square) is rigorously tested.

Chapter five is concerned with the crime database and the crime prediction models. GWPR and GWNBR models are developed, and the models' predictive performance using two bandwidth types (fixed Gaussian and adaptive bi-square) is rigorously tested. The GWNBR model was combined with the empirical Bayes method to identify areas of concern for violent and non-violent crime and to address the issue of regression-to-the-mean.

Chapter six focuses on the application of the models in the identification of crime and collision hotspots in the City of Regina. The empirical Bayes technique is discussed.

Chapter seven provides a summary, a set of conclusions drawn from the research, recommendations, and directions for future research.

Figure 1-1 presents a graphic showing the structure of this dissertation.

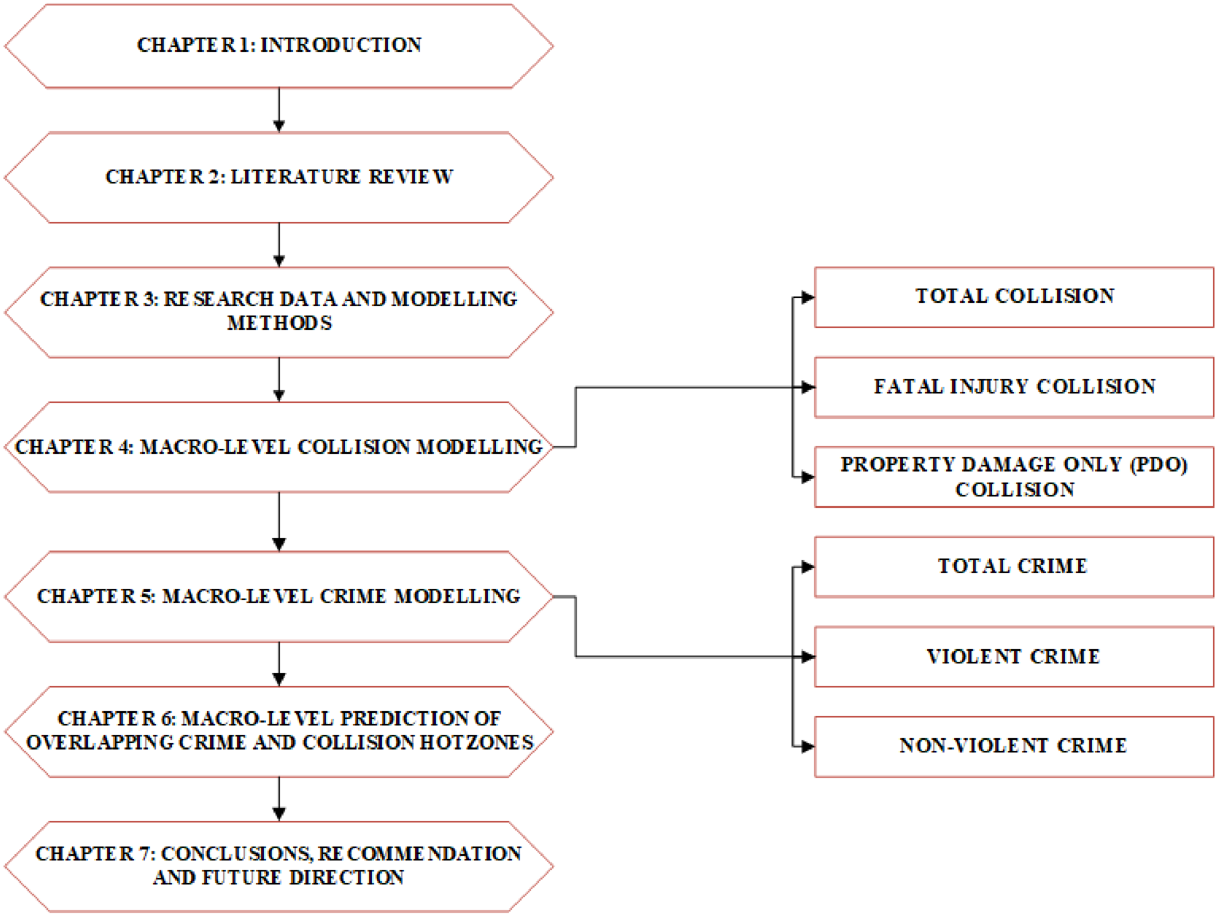

CHAPTER 1: INTRODUCTION
CHAPTER 2: LITERATURE REVIEW
CHAPTER 3: RESEARCH DATA AND MODELLING METHODS
CHAPTER 4: MACRO-LEVEL COLLISION MODELLING
CHAPTER 5: MACRO-LEVEL CRIME MODELLING
CHAPTER 6: MACRO-LEVEL PREDICTION OF OVERLAPPING CRIME AND COLLISION HOTZONES
CHAPTER 7: CONCLUSIONS, RECOMMENDATION AND FUTURE DIRECTION
TOTAL COLLISION
FATAL INJURY COLLISION
PROPERTY DAMAGE ONLY (PDO) COLLISION
TOTAL CRIME
VIOLENT CRIME
NON-VIOLENT CRIME

Figure 2-1: Dissertation Structure

References

Adel, H., Salheen, M., and Mahmoud, R. A. (2016). Crime in relation to urban design. Case study: The Greater Cairo Region. Ain Shams Engineering Journal, 7(3), 925-938.

Alpert, G. P., and Lum, C. (2014). Police pursuits: A complex policy arena. In Police pursuit driving (pp. 1-12). Springer, New York, NY.

Bhagyaiah, M., and Shrinagesh, B. (2014). Traffic analysis and road accidents: a case study of Hyderabad using GIS. In IOP conference series: earth and environmental science (Vol. 20, No. 1, p. 012026). IOP Publishing.

Burch, J., and Geraci, M. (2009). Data-driven approaches to crime and traffic safety. The Police Chief, 76(8).

Carter, J. G., and Piza, E. L. (2017). Spatiotemporal convergence of crime and vehicle crash hotspots: Additional consideration for policing places. Crime and Delinquency, 0011128717714793.

Castillo-Manzano, J. I., Castro-Nuño, M., and Fageda, X., (2015). Are traffic violators criminals? Searching for answers in the experiences of European countries. Transport policy, 38, 86-94.

Coldren Jr, J. R., Huntoon, A., and Medaris, M. (2013). Introducing smart policing: Foundations, principles, and practice. Police Quarterly, 16(3), 275-286.

Cook, C. (2012). Implementation of an area traffic officer program.

Coppola, M. (2014). Policing Model Seeks To Reduce Traffic Crashes and Crime. TechBeat Dated, 3.

Dong, N., Huang, H., Lee, J., Gao, M., and Abdel-Aty, M. (2016). Macroscopic hotspots identification: a Bayesian spatio-temporal interaction approach. Accident Analysis and Prevention, 92, 256-264.

Easton, S., Furness, H., and Brantingham, P. (2014). The cost of crime in Canada. Fraser Institute.

Eck, J., and Spelman, W. (2016). Problem oriented policing. Washington, DC.

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

Gill, C., Weisburd, D., Telep, C. W., Vitter, Z., and Bennett, T. (2014). Community-oriented policing to reduce crime, disorder and fear and increase satisfaction and legitimacy among citizens: a systematic review. Journal of Experimental Criminology, 10(4), 399-428.

Hadayeghi, A., Shalaby, A. S., and Persaud, B. N. (2010). Development of planning level transportation safety tools using geographically weighted Poisson regression. Accident Analysis and Prevention, 42(2), 676-688.

Highway Safety Manual (2010). American Association of State Highway and Transportation Officials. ISBN: 9781560514770

Kamruzzaman, M., Haque, M. M., Ahmed, B., and Yasmin, T. (2013). Analysis of traffic injury severity in a mega city of a developing country.

Keighley, K. (2017). Police-reported crime statistics in Canada, 2016. Juristat: Canadian Centre for Justice Statistics, 3.

Kim, S., and Ulfarsson, G. F. (2008). Curbing automobile use for sustainable transportation: analysis of mode choice on short home-based trips. Transportation, 35(6), 723-737.

Kloog, I., Haim, A., and Portnov, B. A. (2009). Using kernel density function as an urban analysis tool: Investigating the association between nightlight exposure and the incidence of breast cancer in Haifa, Israel. Computers, Environment and Urban Systems, 33(1), 55-63.

Kuo, P. F., Zeng, X., and Lord, D. (2011, November). Guidelines for choosing hot-spot analysis tools based on data characteristics, network restrictions, and time distributions. In Proceedings of the 91 Annual Meeting of the Transportation Research Board (pp. 22-26).

Leigh, J., Dunnett, S., and Jackson, L. (2017). Predictive police patrolling to target hotspots and cover response demand. Annals of Operations Research, 1-16.

Li, Z., Wang, W., Liu, P., Bigham, J. M., and Ragland, D. R. (2013). Using geographically weighted Poisson regression for county-level crash modeling in California. Safety science, 58, 89-97.

Mayhew, D. R., Singhal, D., Simpson, H. M., and Beirness, D. J. (2004). Deaths and injuries to young Canadians from road crashes. Ottawa, Ontario: Traffic Injury Research Foundation.

McCollister, K. E., French, M. T., and Fang, H. (2010). The cost of crime to society: New crime-specific estimates for policy and program evaluation. Drug and alcohol dependence, 108(1), 98-109.

Moon, T. H., Heo, S. Y., and Lee, S. H. (2014). Ubiquitous crime prevention system (UCPS) for a safer city. Procedia Environmental Sciences, 22, 288-301.

National Highway Traffic Safety Administration (NHTSA). (2009). Data-driven approaches to crime and traffic safety (DDACTS): operational guidelines. Washington, DC: US Department of Transportation.

Ratcliffe, J. H. (2004). The hotspot matrix: A framework for the spatio-temporal targeting of crime reduction. Police practice and research, 5(1), 5-23.

Reisig, M. D. (2010). Community and problem-oriented policing. Crime and justice, 39(1), 1-53.

Rydberg, J., McGarrell, E. F., and Norris, A. (2014). Flint DDACTS pilot evaluation. East Lansing, MI: Michigan Justice Statistics Center, School of Criminal Justice, Michigan State University.

Siddiqui, C., Abdel-Aty, M., and Huang, H. (2012). Aggregate nonparametric safety analysis of traffic zones. Accident Analysis and Prevention, 45, 317-325.

Soh, M. B. C. (2012). Crime and urbanization: Revisited Malaysian case. Procedia-Social and Behavioral Sciences, 42, 291-299.

Statistics Canada (2018). Deaths causes of death and life expectancy, 2016. Catalogue no. 11-001-X. https://www150.statcan.gc.ca/n1/daily-quotidien/180628/dq180628b-eng.htm

Stuster, J., Worden, R., McLean, S., and Stuster, J. D. (2010). Data-Driven Approaches to Crime and Traffic Safety (DDACTS): Case studies of six programs.

Sweeney, E. M. (2009, Spring). The value of traffic enforcement. Big Ideas for Smaller Police Departments, 4(2) 1-5.

Takyi, E. A., Oluwajana, S. D., and Park, P. Y. (2018). Development of macro-level crime and collision prediction models to support Data-Driven Approach to Crime and Traffic Safety (DDACTS). Transportation Research Record, 0361198118777356.

Transport Canada (2016). Canadian motor vehicle traffic collision statistics: 2016. Catalogue No. T45-3E-PDF, ISBN 1701-6223. http://www.tc.gc.ca/eng/motorvehiclesafety/canadian-motor-vehicle-traffic-collision-statistics-2016.html

Vemulapalli, S. S., Ulak, M. B., Ozguven, E. E., Sando, T., Horner, M. W., Abdelrazig, Y., & Moses, R. (2017). GIS-based spatial and temporal analysis of aging-involved accidents: a Case Study of Three Counties in Florida. Applied Spatial Analysis and Policy, 10(4), 537-563.

Weisburd, D., Telep, C. W., Hinkle, J. C., and Eck, J. E. (2010). Is problem-oriented policing effective in reducing crime and disorder? Criminology and Public Policy, 9(1), 139-172.

Wilson, J. W., and Heinonen, J. A. (2012). Police workforce structures: Cohorts, the economy, and organizational performance. Police Quarterly, 15, 283-307.

World Health Organization (2018). Road traffic injuries. http://www.who.int/news-room/fact-sheets/detail/road-traffic-injuries

World Health Organization. (2015). Global status report on road safety 2015. World Health Organization.

Xie, Z., and Yan, J. (2008). Kernel density estimation of traffic accidents in a network space. Computers, Environment and Urban Systems, 32(5), 396-406.

Xu, P., Huang, H., Dong, N., and Abdel-Aty, M. (2014). Sensitivity analysis in the context of regional safety modeling: identifying and assessing the modifiable areal unit problem. Accident Analysis and Prevention, 70, 110-120.

Yu, H., Liu, P., Chen, J., and Wang, H. (2014). Comparative analysis of the spatial analysis methods for hotspot identification. Accident Analysis and Prevention, 66, 80-88.

CHAPTER 2:LITERATURE REVIEW

This chapter begins by discussing DDACTS guidelines and principles. It evaluates two perspectives, micro-level and macro-level, to determine areas with a high incidence of crime and collisions. The various units available for macro-level analysis are highlighted and the popularly used units among transportation engineers are emphasised. This chapter also reviews technical challenges associated with developing macro-level crime and collision prediction models, such as the problem of spatial dependency. Regression-to-the-mean problems, which can lead to misidentification of peak periods or seasons for high crime and collisions and inaccurate identification of peak locations (due to random variation) if not accounted for, are discussed. The chapter also discusses modelling techniques such as global models (e.g., Poisson, Poisson Log-normal, and negative binomial) that use fixed parameters to describe relationships across geographic space. The main limitations of the global regression model, e.g., not accounting for spatial dependency, are emphasised. The spatial models for modelling spatial dependency are discussed, but the focus is placed on the geographically weighted regression models because these models are suitable for count data and can account for spatial dependency. Also discussed is the importance of bandwidth to the geographically weighted regression model. Two of the most advanced techniques for adjusting for the regression-to-the-mean bias problem in the identification of hotspots or hotzones are discussed and thoroughly compared in this chapter.

2.1 Data Driven Approach to Crimes and Traffic Safety (DDACTS)

Nowadays, many enforcement authorities are having trouble combating the problem of traffic collisions and criminal activities due to increasing budget cuts (Maggard and Lung, 2009). For example, enforcement strength (measured as the number of police per 100,000 population) in Canada decreased by 1% in 2017 marking six years of decline and the lowest rate since 2004

(Conor, 2018). The decrease can be attributed to budget cuts in enforcement resources across different provinces. For instance, Regina, the capital city of the province of Saskatchewan, had a low enforcement operating budget in 2016. About 89% of the approximately $81 million total operating budget was allocated to wages of service officers and only 8 more officers were added (City of Regina, 2016). Despite the demand for operations continuing to increase, the number of enforcement personnel is not increasing and may not ensure the safety of communities in Regina. Thus, a proactive approach is needed to effectively use available personnel in combating crime and collision occurrences. The popularly sought approach is DDACTS.

DDACTS is a law enforcement tactic and strategy developed collaboratively between the United States Institute of Justice (NIJ), Bureau of Justice (BIJ) and National Highway and Transportation Safety Administration (NHTSA). It aims at the simultaneous reduction of crime and collisions (Burch and Geraci 2009; Cook 2012; Lopez et al., 2018; McGarrell et al., 2014; NHTSA, 2009). DDACTS incorporates crime and collision locational data to establish the appropriate strategy for deploying resources using spatial analysis (kernel density) to determine high crime and collision areas. Through high visibility of enforcement in target areas, enforcement activities aimed at crime and collisions reduction are increasingly guided to make a more significant impact (Carrick et al., 2015; Lopez et al., 2018).

DDACTS focuses on crime and collision data integration for effective planning of enforcement operations (Chen et al., 2016). It provides a solution to the struggle of enforcement authorities with limited budgets to respond to high demand for service in a fashion that address the problems of both crime and collisions (Cook, 2012; Maistros, and Schneider IV, 2017). It helps in maximizing the strained budget of law enforcement through a focus on the most reliable target(s) (minimized geographic focus) and the deployment of resources to locations or areas of priority

(Haberman 2017; Weisburd, 2008). Through intensive enforcement on focus areas, crime and collisions are reduced leading to improved accountability and more effective use of resources (Cook, 2012).

The DDACTS approach considers the social menace of crime as primary while traffic safety is secondary. The current DDACTS approach used by enforcement relies on density estimation for hotspot mapping. It requires separate hotspots for crime and collisions to be created. Through superimposition of these hotspot maps, priority locations (defined as locations or areas with a concurrent high incidence of both crime and collisions) are selected for law enforcement deployment.

According to McClure et al. (2014), DDACTS is founded on principles and guidelines that facilitate its deployment. Figure 2-1 shows the seven DDACTS guiding principles that are designed to facilitate the effective deployment of limited resources.

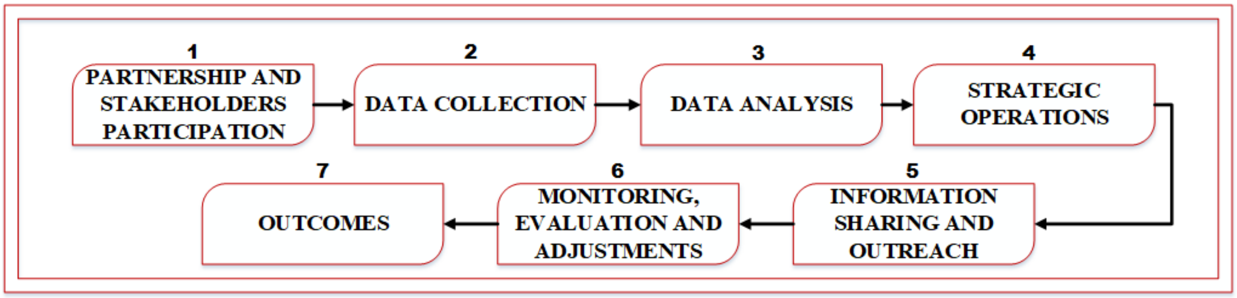

1
PARTNERSHIP AND STAKEHOLDERS PARTICIPATION
2
DATA COLLECTION
3
DATA ANALYSIS
4
STRATEGIC OPERATIONS
5
INFORMATION SHARING AND OUTREACH
6
MONITORING, EVALUATION AND ADJUSTMENTS
7
OUTCOMES

Figure 2-1: DDACTS Guiding Principles

(Adapted from NHTSA, 2009)

The DDACTS principles are:

Partnership and stakeholder participation: Collaboration between community agencies and stakeholders is critical to DDACTS analysis. Through the participation of stakeholders, easy access is granted to data such as crime and collision data that are often collected by diverse agencies within the same jurisdiction. Collaboration enables analysis of data to identify problem areas. Also, stakeholder collaboration helps promote DDACTS implementation. It also provides grounds for the assessment of the impact of DDACTS through feedback on the potential impact of DDACTS in improving public safety.

Data collection: DDACTS requires the collection of data such as confidential crime details, police authority crime classifications, and codes of enforcement within a participating jurisdiction. Traffic collision details of location(s) and person(s) involved are also collected. Information about the causative factor at the micro-level and location coordinates of where crimes and collisions occur are also collected.

Data analysis: Analysis of data provides insights into problem areas where jurisdictions or cities can develop the appropriate strategies to reduce the occurrence of crime and collisions. Problem areas in DDACTS enforcement refer to the areas referred to as hotspot (hotzone in macro-level analysis) where high numbers of crimes and collisions occur. The analysis usually used to determine these areas relies on GIS density estimation techniques.

Strategic operations: This principle uses the data analysis findings to develop appropriate enforcement strategies that could be deployed in minimizing the social harms (loss of life and property, injury) associated with crime and collisions. Strategies deployed to achieve a simultaneous reduction in crime and collisions usually involve high visibility patrols in target areas. The strategic operation principle of DDACTS allows organization changes and facilitates incorporation DDACTS into enforcement tactics rather than relying on the traditional method of random patrols.

Information sharing and outreach: This facilitates adequate communication between stakeholders involved in the deployment of DDACTS within a specific jurisdiction in order to assess the performance of DDACTS in reducing crimes and collisions. It also facilitates obtaining feedback from the community about the results of the implementation of the enforcement tactics.

Monitoring, evaluation and adjustment: This principle allows for jurisdictions implementing DDACTS methodology to continually evaluate the performance of the enforcement model towards achieving the objectives of crime and collision reduction. Through continuous evaluation of the performance of DDACTS, jurisdictions can make appropriate adjustments to improve the outcomes of DDACTS deployment.

Outcome: This principle involves documentation of the result of DDACTS deployment within a specific jurisdiction in terms of crime and collision reduction, and community safety improvements. This principle also allows reporting the influence of DDACTS deployment and communicating and sharing information among participating stakeholders and community affiliates.

In summary, DDACTS principles allow collaboration between the enforcement authority and community stakeholders using available data to identify places that have a high number of crimes and collisions and require strategic operations. An example of a DDACTS hotspot map can be seen in Figure 2-2. In Figure 2-2, the red and the blue colours represent areas of high incidence of crime and collisions respectively. Where these colours overlap are the DDACTS target areas for enforcement deployment where high visibility of enforcement personnel could bring about a simultaneous reduction in crimes and collisions.

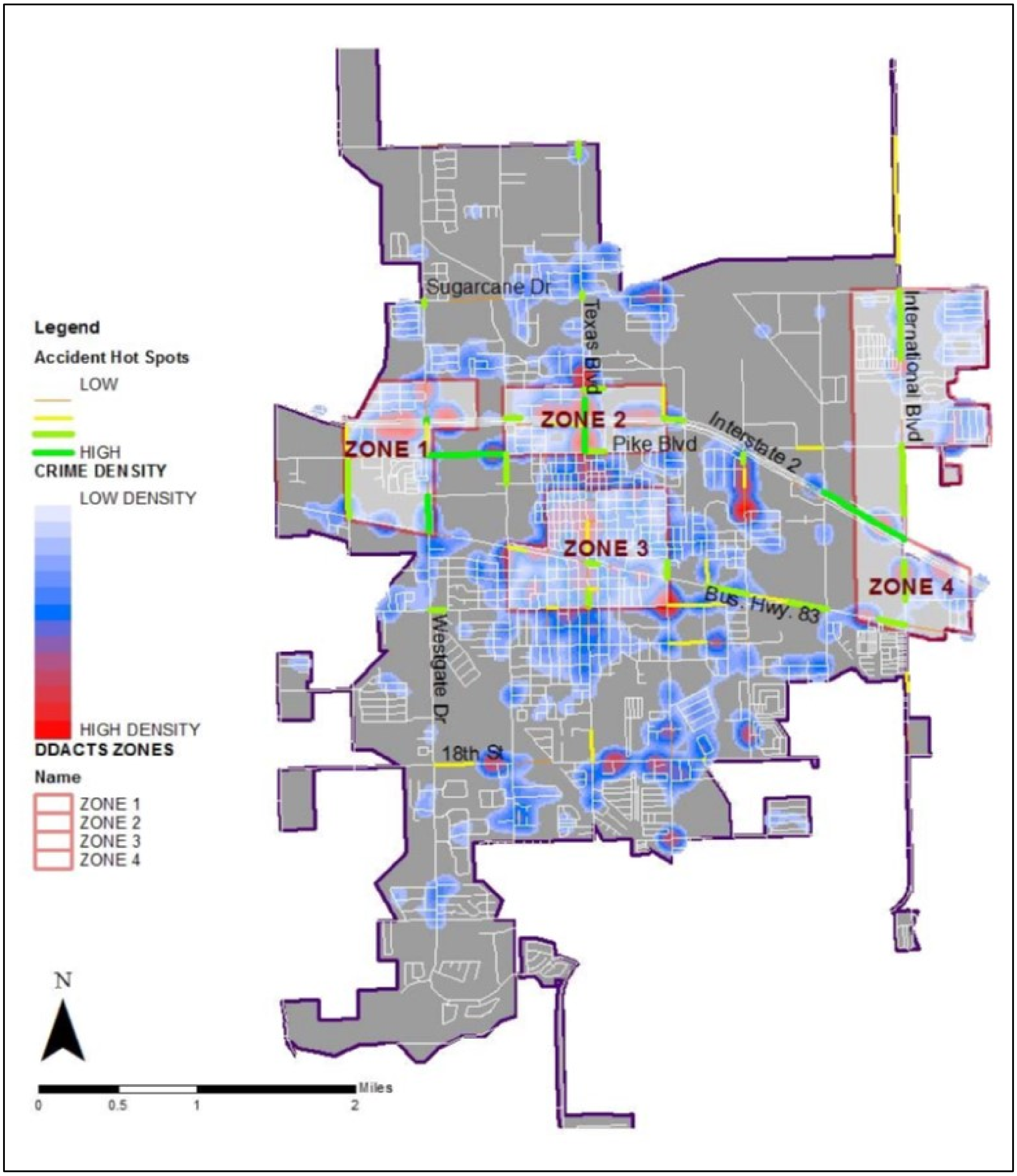


Figure 2-2: The DDACTS Hotspot Map

(Source: Weslaco Police Department, Texas USA)

DDACTS methodology has received great interest from enforcement agencies due to the significant reductions in crime and collisions that have been recorded where DDACTS is applied (McClure, 2014; McGarrell et al., 2014). It has also brought about effective utilization of resources through the efficient operation by enforcement authorities. While this approach has been widely acclaimed as successful in reducing crime and collisions, it does not reveal information (such as demographic, socio-economic and land use details) about the characteristics of hotspots.

In a recent evaluation of DDACTS enforcement tactics, McClure et al. (2014) suggested that environmental factors should be included in hotspot identification in the DDACTS analysis. Incorporation of these factors would facilitate future hotspots prediction such that the DDACTS models could be used for long range planning and policy development. However, environmental factors such as demographic, socio-economic and land use factors that influence crime and collision occurrence are always presented at macro-level, i.e., in aggregated form, to protect privacy. Thus, to incorporate these aggregated factors, hotspot determination must shift to macro-level analysis such that areas with a high incidence of crime can be determined. Recent advances in the DDACTS concept have led to the evaluation of new approaches to hotspot determination to support enforcement operations.

For example, Drawve et al. (2018) explored how risk terrain modelling combined with conjunctive analysis of case configuration could be incorporated into hotspot detection. They used Green Bay, Wisconsin as a case study. The risk terrain model was used to determine the impact of the built environment on the potential for traffic collisions. Risk terrain modelling identifies hotspots by considering the characteristics of the physical or built environment that contributes to the risk. They found 6 risk factors that lead to vehicle collisions: stop signs, retail and clothing stores, bus stops, pedestrian signs, fast-food restaurants, and convenience stores. Conjunctive case configuration was used to determine the environmental configurations that could likely contribute to traffic collisions at priority locations. It determines whether a risk factor is present in all cases of hotspots identified. With the six environmental factors that give rise to traffic collisions identified, about 64 different combinations of these risk factors are implied. Patterns identified by conjunctive case analysis showed 8 leading configurations out of which 4 combinations were found to be dominant. Common combinations of these factors at hotspots identified are the presence of a bus stop, stop sign and fast-food restaurant.

Drawve et al. (2018) concluded that their findings supported the idea that the concept of risky places within an environment could be incorporated into a vulnerability exposure framework which determines factors that gives rise to the incidence of collisions. They opined that the combination of risk terrain modelling and conjunctive case analysis could further help provide guidance to enforcement in identifying locations or areas where future collisions could occur. While this approach seems like an advancement to the DDACTS methods, it has only been evaluated on traffic collisions. The feasibility of the approach for crime hotspot detection has not been evaluated.

Drawve et al. (2018) explained that risk terrain modelling is similar to density estimation techniques but combining it with conjunctive case configuration enhances hot zone determination accuracy. It could, however, be argued that this approach is not predictive. It only identifies micro-level causative independent variables that lead to the occurrence of a collision hotspot. Predictor factors such as demographic, land use and socio-economic factors that exist at the macro-level and give rise to the hotspots are not considered.

To incorporate these factors into hotspot analysis, Takyi et al. (2018) conducted macro-level analysis to incorporate demographic, land use and socio-economic factors in hotspot determination. As the researchers conducted their analysis at a macro level, the hotspots identified are referred to as hotzones.

Takyi et al. (2018) used the City of Regina as a case study and developed separate prediction models for violent crime and fatal-injury collisions. The models considered demographic, land use and road inventory variables. The negative binomial regression model was used. Using the models developed, hotzones for crimes and collision were predicted and DDACTS hotzones determined spatially for enforcement prioritization. The approach used by Takyi et al. serves as an advancement to the DDACTS approach as it is predictive in nature, but it does not consider the spatial dependency (variation in relationship across geographic space) that characterizes spatial data such as crime and collisions. These issues present opportunity for the research discussed in this dissertation.

2.2 Macro-level Modelling of Crimes and Collision

In general, safety analysis of a place, location or area is usually done using one of two approaches: micro level analysis or macro level analysis.

Before discussing macro level modelling, it is worth considering micro level analysis. Micro level analysis focuses on specific crimes and collisions at the individual level (Hautzinger et al., 2007) without much detail about the aggregated the environmental factors that give rise to the prevalence of the crime and collisions.

In transportation engineering, microscopic level analysis focuses on collision incidences at the level of individual road segments or intersections in an attempt to understand how influencing micro-level factors (such as the geometric design, lighting and traffic flow characteristics) affect collision occurrence (Cai, 2017; Hautzinger et al., 2007; Huang et al., 2016). Micro-level approaches to modelling collisions have been extensively used in transportation engineering (Abdel-Aty and Radwan, 2000; Huang et al. 2017; Hauer et al., 2004; Kononov et al., 2008; Pulugurtha and Sambhara, 2011; Wang et al. 2015; Zhang and Ivan, 2005). Micro-level analysis of collisions forms an essential component of the network screening process recommended by the Highway Safety Manual (2010).

Micro-level analysis has also been used in crime studies (Curman et al., 2015; Piza et al., 2014; Rosser et al., 2017; Vandevuver and Steenbeck, 2017). In micro-level crime modelling, specific individuals that commit crimes are targeted. This type of analysis focuses on individual level factors that gives rise to the occurrence of crime (Sporer et al., 2016). In most cases, it explores criminology principles and micro-level theories such as discrete choice to establish the motivations of offenders committing crimes in various different areas. An example is the research conducted by Bernasco (2010) in which the locations chosen by criminals to perpetrate their crimes were examined using discrete choice. Bernasco (2010) emphasised that this type of approach is necessary to determine offenders' motives for selecting a target location. Micro-level crime modelling may also use anonymous surveys to evaluate individual differences in criminal behaviour (Wells et al., 2012).

The difference between macro and micro-level crime or collision studies is the unit of analysis used for the independent variables that influence the occurrences of interest. The level may be individual or aggregated (Sporer et al., 2016). It was emphasised by Sporer et al. (2016) that when policies are to be targeted at individuals to prevent individuals from conducting criminal behaviour, micro-level analysis is suitable.

Macro-level analysis, however, uses large scale (zonal) socio-economic, demographic and land use characteristic for modelling and prediction (Lee et al., 2015). This approach has become a preferred technique that facilitates the development of multivariate models using aggregated information (Xu et al., 2016). Macro-level analysis relates aggregated statistics such as traffic collisions or crimes to a spatial unit with other macro-level variables that exist in the same spatial unit (Huang et al., 2016). When the focus of any analysis is on program and policy formulation for a large area, macro-level analysis and modelling is suitable. In this dissertation, the focus is on an area wide approach to crime and collision reduction.

Macro-level analysis has been extensively used in crime studies (Britt et al., 2018; Boivin and Felson, 2018; Freilich et al., 2015; Kim, 2018; Light and Harris, 2012; Simes, 2017; Steenbeek and Weisburd, 2016; Townsley et al., 2016). It has also found extensive application in collision studies (Alkahtani et al., 2018; Cheung et al., 2008; Hadayeghi et al., 2006; Huang et al., 2010; Lee and Abdel-Aty; 2018; Lee et al., 2018; Ossama and Sayed, 2016; Wang et al., 2016; Zhai et al., 2018).

Evaluating influencing factors such as the land use and socio-economics, etc. that contribute to the prevalence of crimes and collisions at different locations requires analysis at the level of spatial units. Various spatial units of aggregation exist for macro-level analysis and range from regional to block level. The hierarchy of spatial units in Canada is shown in Figure 2-3.

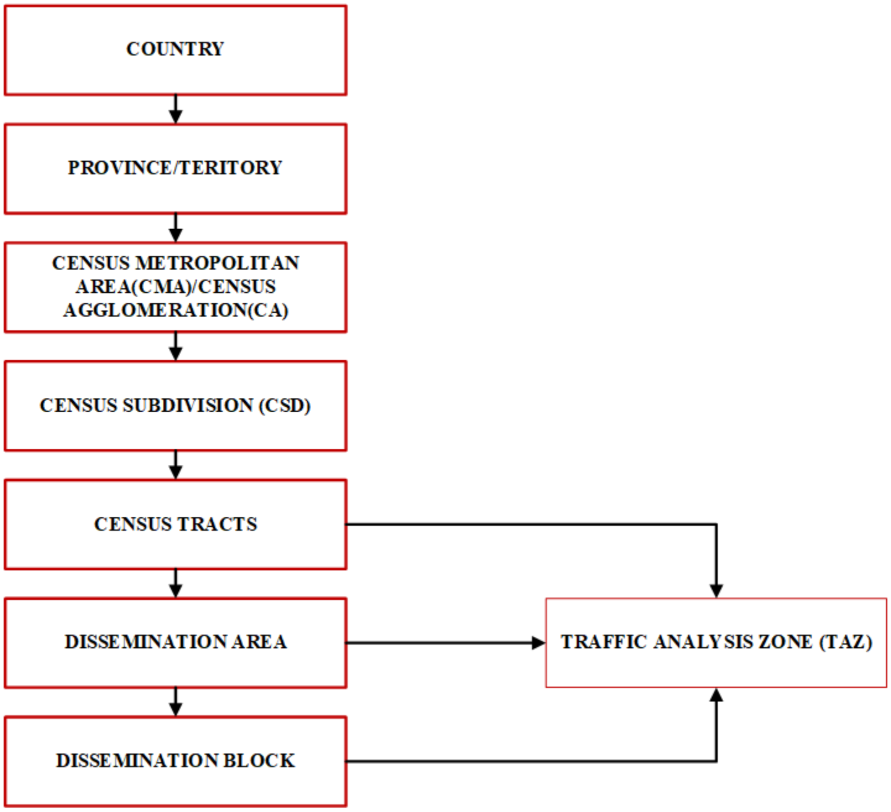


Figure 2-3: The Hierarchy of Census Geography for Information Collection in Canada Statistics Canada, (2011)

Macro-level models strive to establish relationships between variations in the number of incidences (e.g., number of crimes or collisions) and their independent variables (as available at the level of geographic unit selected). The increased acceptance of macro-level modelling has been due to the recognition that factors (e.g., demographics, socio-economics and land use) affecting both crimes and collisions have spatial dimensions. These spatial factors are presented at a high level of aggregation (e.g., zonal) due to privacy issues. Furthermore, the confidentiality imposed on socio-economic, demographic and land use data applies to both crime and collision data and led to the presentation of the information at the zonal level. This has become standard for long range planning (Lee et al., 2015; Xu et al., 2014).

However, analysis at the zonal level is not without its challenges.

Spatial Analysis Units in Macro-Level Analysis

Macro level safety analysis is essential to safety planning and influences decisions, policies and investment with respect to safety (Abdel Aty et al, 2013). In road safety, it has become the new proactive approach for better understanding of the incidence of collisions at area level (Naderan and Shahi, 2010). This also applies to the incidence of crime. Spatial aggregation approaches vary considerably. The choice used in macro-level modelling depends on the researcher's choice and the spatial units at which data for the independent variables that influence crime and collisions are available. The popularly used spatial unit among transportation engineering professionals is the Traffic Analysis Zone.

Traffic Analysis Zone (TAZ)

A Traffic Analysis Zone (TAZ) is a statistical entity used by transportation professionals for data collection. TAZs are spatial analysis units used to determine the movement of people and goods in transportation planning and analysis (Jeon et al, 2012; Metaxatos, 2015). They are mostly used in transportation engineering to represent trip origins and destinations. TAZs contain information about the socio-economic, land use and demographic characteristics of each zone.

TAZs are usually created considering the homogeneity of demographic, land use and socio-economic characteristics and the minimization of intra-zonal trips while recognising physical, political and historical boundaries (Cambridge Systematics and AECOM, 2007, Pulugurtha et al., 2013). TAZs are critical to transportation planning studies due to the large amount of information collected at the TAZ level. The unit has become a key element of transportation demand studies (Martinez et al., 2010). The information collected can be used in zonal level crime analysis.

Since DDACTS analysis of crime and collisions is carried out at a macro-level, the TAZ is the spatial unit of analysis used in this dissertation. The macro-level variables required for the research are available at the TAZ level.

2.3 Technical Challenges in Macro-Level Collision and Crime Prediction Modelling

When data are aggregated into areal units for zonal level analysis, problems of spatial dependency and spatial heterogeneity often arise (Dong et al., 2014; Haining et al., 2009).

Spatial dependency refers to relationships that exist between observation units when the units are in proximity leading to similarities in the value of observations. It implies that, to some extent, observations within proximal locations share some amount of the same information, i.e. systemic patterning of data may be recorded at neighbouring locations (Collins et al., 2006; Paez and Scott, 2005; Plumper and Neumayer, 2010). Spatial dependence in nearby locations leads to violation of the independence of observations assumed by standard statistical modelling techniques (Bernasco and Elffers, 2010; Collins et al., 2006; Fotheringham, 2009; Getis, 2008; Lee, 2017).

Spatial heterogeneity represents variation in relationships that exist between observation units across geographic space (Basile et al, 2014; Getis, 1999; DiRienzo et al., 2000; Li and Zhu, 2015; Zhang et al., 2009). It usually arises as a result of uneven distribution of event across a region.

Spatial dependency and heterogeneity are closely related and distinguishing between the two can often be challenging (Jiang, 2015; Plumper and Neumayer, 2010). They co-exist and are both components of spatial effects (Zhang et al, 2009). Spatial heterogeneity is a first order spatial effect while spatial dependency is a second order spatial effect.

Spatial dependency could lead to spatial heterogeneity. When high values events cluster together, the result is spatial dependency and could lead to uneven distribution of event across space, known as spatial heterogeneity. To simplify the issue, this dissertation uses the term spatial dependency exclusively.

Another challenge that arises in macro-level analysis of crime and collisions for hotzone determination is Regression-to-the-mean (RTM) bias. This refers to a statistical phenomenon of repeated measurements in which random variation appears to be real variation (Barnnett et al., 2004, De Pauw et al., 2013; Elvik et al., 2009).

It important that the technical challenges such as spatial dependency and RTM bias are taken into consideration in any attempt to determine areas with a high incidence of crime and collisions at the macro-level.

2.3.1 Spatial Dependency and Heterogeneity

The problem of spatial dependency and heterogeneity is peculiar to aggregated data used in zonal level analysis. Data in which spatial effects (spatial dependency and heterogeneity) are known to occur include crime and collision data. Collision data are usually collected with reference to location in space (Quddus, 2008). So also are incidences of crime. When the spatial effects are not taken into consideration in modelling, the analysis is described as global models. Global analysis uses fixed parameters in describing relationships that exist across geographic space.

The assumption of a fixed parameter for data from a geographic reference is often incorrect (Haining, 2009; Meng, 2014; Li et al., 2011; Lloyd and Shuttleworth, 2005). Thus, accommodating spatial dependency and heterogeneity in modelling variables with spatial dimension becomes important. This is achieved by incorporating spatial weights for the observation locations.

The spatial clustering seen in macro-level modelling could be attributed to boundaries imposed on continuous geographic surfaces. In most cases, the boundaries used to delineate neighbouring zones are artificial so there is a likelihood of similarity in characteristics in dense areas when examined as a continuous surface.

A common example of spatial clustering of events in collision studies relates to Vehicle Kilometers Travelled (VKMT) which is used as exposure for modelling traffic collisions. It is generally accepted that as VKMT increases, the likelihood of collisions increases (Rolison and Moutari, 2017; Rhee et al., 2016; Soltani and Askari, 2017; Zegrass, 2010). TAZs with higher VKMT often tend to cluster giving an indication of spatial dependencies between nearby locations and thus similarities in the numbers of collisions observed in zones that are close in proximity. VKMT similarities in zones within close proximity could be linked to highly similar traffic characteristic in neighbouring zones. The artificial boundaries-imposed cause spatial clustering.

For examples, high volume arterial roads usually cross proximal zones. Spatial dependency associated with VKMT could also be linked to the conditions used in aggregating the information used to determine the boundaries. In the case of VKMT and zones with an arterial in common, the assumption is that both zones contribute equally to the traffic on the arterial. This means that the boundary that separates the zones could result in the clustering of collisions in both zones. This then gives rise to spatial dependency between zones with a boundary and major arterial in common. It is clear that VKMT and socio-economic and land use characteristics may contribute to the clustering of traffic collisions (Rhee et al., 2016; Soltani and Askari, 2017).

The presence of spatial dependency in crime data has been extensively discussed in crime studies (Deane et al., 2008; Light and Harris, 2012). According to Collins et al. (2006), the clustering of crimes at certain locations and areas within a city or neighbourhood strongly influences characteristics and social interaction within areas. The higher rates of crime (both violent and non-violent) seen in some neighbourhoods could be attributed to high clustering of unemployed persons and/or low-income characteristics of the area. This type of socio-economic characteristic is typically spatially clustered within an area thus contributing to spatial dependency in the crime rate. Also, interactions not confined to an area may cause crime to spill to other adjacent location thereby leading to similarities in crime rates (Haining et al., 2009).

A solution to the challenges of spatial dependency requires allowing model coefficients to vary locally by incorporating additional information based on the spatial structure (Anselin, 2010). Specifying local relationships across space requires the use of spatial regression models and can be classified into three broad categories (Bernasco, and Elffers, 2010). These are the spatial error models, spatial lag models, and the geographically weighted regression models.

According to Anselin (2003), a spatial error model allows for spatial dependency in the error of a model and is dependent on the spatial weight matrix defined for a model. In a spatial error model, the error term is divided into uncorrelated and correlated parts. Bernasco and Elffers (2010) explained that a spatial error model is suitable if there is a possibility of spatial dependency in independent variables affecting a response variable. Spatial error models take into consideration spatial influences in unobserved variables (spatial dependency is specified on error term) (Ward and Gleditsch, 2018). A spatial error model is most suitable where there is a possibility of interdependence in the error term rather than in the independent and dependent variable.

A spatial lag model evaluates spatial interdependency in variables across units of analysis in a geographic space. It uses the observations from proximal locations or areas to provide a reason for the occurrence of an event in nearby areas. Rather than specifying spatial dependency on the error term, spatial dependency is specified on the fixed parameter estimate of the independent variables of the model. A spatial lag regression model is like an ordinary regression model especially when the lag is placed on the independent variables While a spatial lag model allows for spatial dependency to be incorporated into modelling, interpretation of the results is considered more complex (Chi and Zhu, 2008).

Neither the spatial error nor the spatial lag model calibrates a local regression equation for all data points. The models do not give an indication of how the relationship between the dependent and the independent variables varies across space.

Another spatial regression model used in capturing spatial dependency in data is known as the geographically weighted regression model. This was proposed by Brunsdon et al. (1998) and allows model parameters to vary locally with a known parametric family of distributions placed on the response. Geographically weighted regression models (ordinary least square, logistic, Poisson or negative binomial) are calibrated using bandwidth and by weighting observations based on their proximity to reference point such that observations closer to the reference points are given higher weights than observations for more distant locations (see chapter 3).

The main advantage of these models over the spatial error or spatial lag model is that they allow various distributions to be specified on the response variable rather than following the normal distribution. The models also calibrate a separate regression model for each data point and provide opportunity for spatial dependency in relationships to be thoroughly tested.

This dissertation uses and explores the geographically weighted regression model as its spatial model.

2.3.2 Regression-to-the-Mean (RTM)

RTM is a statistical tendency in repeated measurements whereby an extreme period of measurement is followed by a less extreme period of measurement causing random variation in observations (Davis, 2007, FDOT, 2015; Li et al., 2015; Highway Safety Manual, 2010; Zhao et al., 2015). It makes variations that are natural appear like a real change (Barnett et al., 2004).

The concept of RTM implies that extreme values do not last in the long run, i.e., they average out (Autey, 2012; Senn, 2011). It has become recognized in statistics that the effect of RTM can lead to misinterpretation of research findings (Burrell et al., 2010). For example, the effects of treatments/countermeasures introduced to minimize or reduce crime could be considered effective, but could be simply a result of RTM (i.e., the random variation in, for example, the number of crimes or collisions). Not taking RTM into consideration can lead the selection of hotspots that are not stable, thus leading to ineffective utilization of resources to combat crime and collisions (Highway Safety Manual, 2010).

Marchant (2004), for example, noted that RTM that is not accounted for leads to bias in crime studies. This was revealed in the systematic review of the impact of street lighting conditions on crime using 251 previously published research articles. In the study, previous research on the relationship between crime and lighting conditions was evaluated to determine the validity of the claims that effective lighting reduces crime. On examining this relationship, Marchant found that the effect of street light reducing crimes cannot be generalized. The likely reason for the findings of the previous research was RTM bias based on the difference in the number of crimes recorded in newly lit and control areas. When the effect of lighting was examined using a time series plot, lighting showed no significant impact on the number of crimes.

Differences in the characteristics of areas in which crime numbers are observed is another contributing factor to the problem of RTM in crime studies. As criminal activities at locations vary, areas with a high incidence of crime are expected to have low records in the future and vice versa. Thus, comparisons of areas with low and high crimes rates leads to bias (Marchant and Hall, 2007). Usually, when an intervention is applied in an area with a high incidence of crime, the number of crimes is expected to decrease due to natural fluctuation. This is often interpreted as the effect of these treatments. On the other hand, an increase could be seen when such interventions are applied to areas with low crimes. These problems are a result of the systemic error known as RTM that often leads to biased conclusions. RTM can also be misleading about the effect of a treatment applied (Farrington and Welsh, 2006) as it can create an effect where there is no effect.

RTM bias is not peculiar to crime studies. It has also been reported in collision studies that aimed at providing countermeasures to minimize the number of traffic fatalities. A similar random element in collisions has been reported. On some road segments or in some neighbourhoods, the high number of incidences experienced could be a result of chance. Consequently, there are greater likelihoods that a high frequency of collisions at a place might fall in subsequent years. It is important that the problems of randomness and RTM bias be considered in collision studies. Otherwise, the effect of road safety measures could be overestimated (Geurts, and Wets, 2003; Loo and Anderson, 2015).

Yang and Loo (2016) emphasized that RTM bias can cause exaggeration of the effectiveness of measures aimed at reducing collisions if caution is not taken. This issue was also highlighted by Maher and Mountain (2009). It can also lead to the wrong selection of hotspots for countermeasures if not taken into consideration.

RTM bias in a collisions study can also contribute to migration effects. Migration effects refer to the phenomenon whereby an increase in the number of collisions is seen in locations adjacent to a hotspot in which a road safety improvement have been applied. Migration effects have been also reported in crime studies and are commonly referred to as crime displacement (Park et al., 2012). The migration effects of both crimes and collisions from a hotspot to a proximal location has been largely attributed to the RTM effect caused by bias in the selection of sites (Geurts, and Wets, 2003).

A common approach to evaluating road safety and identifying hotspots involves analysis of the frequency of collisions on the road infrastructure of a selected area. The approach uses historical records to rank sites (intersections or road segments). Most authorities focus on the top ten sites that recorded the highest number of collisions. However, using a relatively short time period data to determine the locations to receive safety improvement means that it is possible to miss hotspots and locations that need improvement. This is due to the nature of collisions: they are events that fluctuate over time. A site experiencing a high number of collisions in a particular year may experience a considerable drop in the frequency of events due to RTM effects (Rodegerdts et al., 2004).

After a high number of collisions, a lower number of collisions may well be experienced in the subsequent study or observation period. Such decreases are mostly attributed to the disappearance of the positive randomness, i.e., the observed higher number collisions. Regardless of the safety improvement measures applied to areas with high collision occurrence, a drop in the number of incidences is likely to be observed. On the other hand, the disappearance of negative randomness occurs when locations with a low incidence of collisions experiences a higher number of collisions in subsequent years of measurement (Yang and Loo, 2016).

An example of research that ignored RTM is work undertaken by Datta et al. (2000) in the study of collision frequency at 18 signalized intersections in Detroit city. Through a partnership with private agencies, countermeasures aimed at improving safety at these intersections were provided. A before-and-after observational study was carried out on three of the intersections with safety improvements. The researchers reported that analysis conducted on the three intersections showed that the countermeasures improved safety. Variability in collision occurrences was ignored in the research and not taken into account in the selection of the intersection hotspots, implying that the effect of countermeasures provided could be effective or ineffective on the long-run.

Collision rates at the selected intersections in the study by Datta et al. (2000) were 95% higher than other intersections within the study area. Hauer et al. (2002) showed that the research by Datta et al. and the claim that countermeasures provided at high collision intersections in Detroit city had brought about a reduction in collision occurrences was biased.

In recognition of RTM, the Institute of Transportation Engineers (2009) recommends that methods that account for random variability must be used to correctly determine hotspots and to evaluate the effect of any treatment applied to the transportation infrastructure. This is also emphasised by the Highway Safety Manual (2010).

2.4 Regression Methods in Count Data Modelling

Regression analysis is one of the main statistical tools for understanding functional relationships between dependent variables and the independent variables (Chatterjee and Hadi; 2015; Loo and Anderson, 2015; Mense, 2001; Montgomery et al., 2012). In this research, crimes and collisions are the dependent variables, and factors such as socio-economic, land use and demographic characteristics are the predictors. Regression can be linear or nonlinear, but linear regression methods will be the focus of this research.

Linear regression can be univariate, i.e., use one dependent and predictor variable, or multivariate, i.e., use several predictor variables. Multivariate regression techniques use multiple predictors in an attempt to account for as much of the variation in the dependent variable as possible (Uyanik and Guler, 2013). The techniques summarize changes in dependent variables as a function of other variables. A popular type of regression is generalized linear regression, and this is the approach adopted by this research.

Generalized linear regression allows exponential families of distribution and has been widely used to understand event occurrences across space. At the microscopic level, it has been widely used in crime and collision prediction. Most importantly in transportation engineering, it has formed the basis on which most hotspot determination techniques are based. For example, generalized linear regression is used to derive safety performance functions (SPFs) which model the relationship between collisions (on segments or at intersections) and average annual daily traffic, segment length and various characteristics of the segment or intersection. Generalized linear regression is the basis for the network screening technique that is prescribed by the Highway Safety Manual (2010) and widely applied in the identification of hotspots.

Manan et al. (2013) used generalized linear (negative binomial) models to predict hotspots for motorcycle collisions on primary roads in Malaysia. The study used a negative binomial to predict fatal motorcycle collision locations on Malaysian primary roads. The results indicated that there was a positive association between motorcycle fatalities and the average daily number of motorcycles and number of access points per kilometer.

Giuffre et al (2014) estimated the safety performance of one way four leg un-signalized intersections in the City of Palermo, Italy using 92 intersections. The data used in the analysis were for 2006 to 2013. Geometric design and traffic flow characteristics were used as predictors in the development of the models. The findings showed that the negative binomial model allows flexibility in capturing variance and outperforms other models. It was, however, suggested that for correlation within the responses to be captured, a Generalized Estimating Equation (GEE) could be used.

The above studies highlight the importance of regression methods in safety studies.

Regression methods have also been widely applied in long range safety planning to understand factors predicting collisions at zonal level. The research includes numerous applications of the regression method to improve understanding of crime events across space. In general, most crime and collision studies have focused on the use of generalized linear regression methods. Different types of generalized linear regression models can be used. The next section focuses on count data models relevant to this research: generalized linear regression using include Poisson, negative binomial, Poisson log-normal, and the zero inflated versions.

In general, these models can be grouped into global models and local models with varying regression parameter. Global models assume that the association between independent and dependent variable over geographic space is constant, i.e., observations are independent of each other (Fotheringham et al., 2002) These types of model are often referred to as traditional regression methods and assumes that the parameters of a model remained the same for the entire study area. Section 2.4.1 discusses global regression models using Poisson regression, Poisson log-normal regression, negative binomial regression, and zero inflated regression.

2.4.1 Global Regression Model for Count Data

Poisson Regression Models

Count data refer to non-negative and discrete events that usually occur within a certain time interval. Count event occurrences do not usually follow a normal distribution. Since events occurs within a certain time interval, they can be described as a rate, and discrete distributions recognizing this property of count data are suitable. Thus, models with Poisson error distribution (e.g., Poisson regression) are most often used (Fahrmeir et al., 2013). Poisson error distribution is considered the starting reference for count data modelling (Walters, 2007; Winkelman, 2008). According to Winkelman, (2008), Poisson regression method is simple and robust when estimating the mean parameters in a log linear relation. However, there are cases where Poisson regression methods are unsuitable. These are situations when it is a requirement that the characteristics of a population other than the mean, such as the variances, be estimated. Importantly, Poisson regression is more suitable when it is a requirement that the mean be modelled. In general, Poisson regression methods establish relationships between the probability density function of a dependent variable and the predictor variables.

Consider y_i, a dependent variable (crimes or collisions) related to sets of predictors x_i. The standard form of Poisson regression method is given as the conditional probability as:

$$P(y|\mu) \ = \frac{exp(-\mu_i)\mu^y}{y_i!} \qquad \textit{(2-1)}$$

where:

μ_i is the mean of the expected value of y_i.

The functional form of the Poisson regression model representing this relationship is given as Equations (2-2) and (3-3):

$$ln(\mu_i) = x_i^T \beta + \varepsilon_i \tag{2-2}$$

$$\mu_i = exp(x_i^T \beta + \varepsilon_i\,) \tag{2-3}$$

$$y_i \sim x_i^T \beta \tag{2-4}$$

where:

y_i is the observed value of the dependent variable for location i;

x_i is the independent variable for location i; and

β is the coefficient for the independent variable.

By combining Equations (2-1) and (2-4), Equation (2-5) is obtained:

$$P(y_i|x) \;= \frac{exp\left(-exp(x_i^T \beta)\right) exp(y_i x_i^T \beta)}{y_i!} \tag{2-5}$$

A Poisson regression method determines the conditional mean $E(y|x)$ and the variance mean $Var(y|x)$ using a single parameter μ expressed as Equation (2-3). This gives the equal dispersion characteristics of Poisson regression:

$$E(y|x) = \mu = Var(y|x)$$

The likelihood function of Poisson regression model is given as 2-6 and the log likelihood for parameter estimation is given as 2-7:

$$L(\beta|y_i) \;= \prod_{i=1}^{n} \frac{exp\left(-exp(x_i^T \beta)\right) exp(y_i x_i^T \beta)}{y_i!} \tag{2-6}$$

$$ln(\beta|y_i) = -\sum_{i=1}^{n} exp(x_i^T\beta) + \sum_{i=1}^{n} y_i x_i^T\beta - \sum_{i=1}^{n} ln(y_i!) \qquad \textit{(2-7)}$$

The Poisson regression model has been used extensively in crime studies. For example, Dieffenbach and West (2001) examined the relationship between people's perception about crimes and the crime rate. They explained that while relationships between television exposure and violent crimes have been thoroughly studied, the effect of television exposure on the perception of property crimes has not been examined. Using Poisson regression, Diefenbach and West (2001) found that persons more exposed to television (compared with people who spent less time watching television) were more likely to have the opinion that violent crime occurs at higher rates and property crimes at a lesser rate. Their research emphasizes the influence of television news. Poisson regression was found useful in relating perception of crime rates to exposure to television.

Similarly, Kondo et al. (2016) used Poisson regression to understand how green space and vacant space influence crime. By focusing on a program of green initiatives that aimed to reuse green space, analysis was performed to determine the influence of the program on crime. This was compared with randomly selected areas that were not the focus of reuse programs. The effects of treating vacant lots with greening stabilization or reuse were compared. They found that greening stabilization of lots reduced the number of burglaries in Youngstown, Ohio. However, a significant increase in vehicle theft crimes was recorded for both treatments. Kondo et al. emphasized that greening vacant lots could bring about a significant reduction in violent crime. While negative binomial models are known to be suitable for over-dispersed count crime data, Kondo et al. claimed the results of both the Poisson and the negative binomial model were the same. Over-dispersion was downplayed.

Further, Poisson regression was used by Beland and Brent (2018) to establish a relationship between crime frequency and traffic congestion in the City of Los Angeles, California. Beland and Brent found that an increase in traffic congestion could be related to increased domestic violence due to the stress experienced by drivers in traffic congestion. Traffic congestion affects emotions and is a cost on human psychology.

Crime research continues to use Poisson models for count crime data modelling owing to the fact that the majority of research in this area is focused on exploratory rather than predictive models.

Application of the Poisson regression method is not limited to crimes studies only. It has been applied in collision studies. Most of the researches that have used Poisson regression in collision prediction were done before 2000. Examples of this research include research conducted by Miaou et al. (1992) to establish an association between truck collisions and roadway geometric features. Similar application of Poisson regression method in modelling collisions could be seen in Maher and Summersgill (1996).

Since early 2000, the attention of transportation engineers has shifted from the use of Poisson models for count collision data due to the assumption of equi-dispersion, i.e. equal mean and variance, which is not usually met. In most cases, the variance is greater than the mean for count data and this phenomenon is described as over-dispersion. The increasing recognition of the limitations of Poisson regression method in modelling collisions have favoured the use of negative binomial models for count data. Negative binomial models are nowadays accepted as the baseline for modelling collision data.

Negative Binomial Regression Models

A negative binomial model is generally used when the Poisson assumption that the mean equals the variance is not satisfied. In situations where this condition is not met, the data could be under-dispersed, and the parameter estimates biased due to inflation of the model variance. Under-dispersion usually occurs when the mean is less that the variance. On the other hand, over-dispersion implies greater variance. Over- dispersion usually arises in data due to different reasons. However, a common reason for over- dispersion is related to the omission of variables that influence a process across observations. When data are characterized by over-dispersion, a negative binomial error distribution model is preferred (Washington et al., 2003).

The functional form of the negative binomial model can be expressed similarly to Equations (2-2) and (2-3). In this case y_i, follows a negative binomial distribution. The probability density function for the negative binomial model can be written as Equation (2-8) with dispersion parameter α:

$$P(y_i|\mu_{i,}\alpha) = \frac{\Gamma\left(\frac{1}{\alpha}+y_i\right)}{\Gamma(y_i+1)\Gamma\left(\frac{1}{\alpha}\right)}\left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}}\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} \tag{2-8}$$

The likelihood function of Equation (2-8) is:

$$L(\mu|y,\alpha) = \prod_{i=1}^{n} exp\left\{y_i \, ln\left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right) - \frac{1}{\alpha}ln(1+\alpha\mu_i) + ln\,\Gamma\left(\frac{1}{\alpha}+y_i\right) - ln\,\Gamma(y_i+1) - ln\,\Gamma\left(\frac{1}{\alpha}\right)\right\} \tag{2-9}$$

The negative binomial regression model can be written using the expected mean of the response shown in Equation (2-2). Thus, the log likelihood function of negative binomial model is given in Equation (2-10).

$$ln(\beta|y,\alpha) = \sum_{i=1}^{n} y_i ln\left(\frac{\alpha exp(x_i^T\beta)}{1+\alpha exp(x_i^T\beta)}\right) - \frac{1}{\alpha} ln\left(1+\alpha exp(x_i^T\beta)\right) + ln\Gamma\left(\frac{1}{\alpha}+y_i\right) - ln\Gamma(y_i+1) - ln\Gamma\left(\frac{1}{\alpha}\right) \quad \textit{(2-10)}$$

Negative binomial regression has been applied in crimes studies. Lo and Zhong (2006) explored the effect of marital status (single or divorced) by gender on arrests for a serious crime using negative binomial regression at the macro-level. They evaluated the hypothesis that different factors influence the criminal tendencies of men and women regardless of the characteristics of the neighbourhood or the environment where they lived in Ohio. Their research suggested that being married or divorced had a greater effect on women's criminal tendencies than on men's.

Similarly, the effect of foreclosure on crime frequency has been investigated using negative binomial regression. Ellen et al. (2013) studied the impact of mortgage foreclosure on criminal activities within a neighbourhood. They used information on crime and foreclosure in New York to determine whether there was a relationship. Results obtained from their studies showed that as the foreclosed property rate increased in a neighbourhood, the crime rate increased. This effect was found to be significantly larger when foreclosure was measured in terms of properties that were about to be auctioned or taken over by banks. Findings similar to those of Ellen et al. (2013) have been reported by Stucky et al. (2012).

Further, Lee and Thomas (2010) studied the effect of changes in population on violent crime in counties of rural communities. They hypothesized that communities that are more likely to work together to protect public values are likely to experience reduced violent crime rates and that the crime rate would be unlikely to increase. Another hypothesis considered was the idea that continual changes in population size will lead to more crime being committed. Results from negative binomial regression showed their hypotheses to be correct. They found that communities that work together to protect public values can withstand the effect of changes in population size on crime, but continual changes in population (population growth) may weaken this effect.

Haberman and Ratcliffe (2015) investigated locations with a high robbery rate at a census block and street level in Philadelphia, Pennsylvania from 2009 to 2011. This was done to determine whether crime location vary across space with time. Within the census block, the number of robbery crimes was observed for four-time intervals and modelled using negative binomial regression. Criminogenic factors such as the market for illegal goods were considered as predictors for robbery. The findings suggested that some places within a census block attracted a higher number of robberies depending on the time of the day.

Pridemore and Grubesic (2012) studied how social organization (e.g., communication, cohesion, and pattern of relationship between societal groups) could be used to better understand the relationship between violent (assault) crime and alcohol outlet density at neighbourhood level. They hypothesized that the relationship between violent crime and alcohol outlets could better be explained using social organization structure. By using 298 block groups in Cincinnati, the effect of social organization was tested with negative binomial regression. They found that the effect of alcohol outlet density on crime was relatively weak in a socially organized environment. In contrast, the effect of alcohol outlet density was predominant and strong in socially disorganized areas.

The application of negative binomial regression has not been limited to crime studies. It has been widely applied in the area of transportation safety to identify countermeasures to reduce the number of fatalities. It is recommended by the Highway Safety Manual (2010) and has been widely referred to as the state of the industry model in road safety.

To mention a few applications, Lovegrove et al. (2010) applied negative binomial regression in the development of macro-level (TAZ) collision prediction models. The research focused primarily on testing safety guidelines in long range transportation safety planning. Using data for 400 neighbourhoods in greater Vancouver, British Columbia, a regression model was calibrated, and the expected mean collision predicted. It was emphasized that the use of negative binomial regression in the development of a zonal level collision prediction model offers safety planners and engineers numerous advantages as factors that influence collisions could be determined at the neighbourhood level.

Wei and Lovegrove (2013) used negative binomial regression in the development of prediction models for cyclist collisions at a zonal level. Focusing on TAZs as the unit of analysis, they used collision records involving bicyclists and vehicles in the Central Okanagan Regional District of British Columbia. The study results revealed that total lane kilometers, bicycle lane kilometers, and the density of bus stops, intersections, traffic signals, and arterial-local intersections significantly influenced the occurrence of cyclist-vehicle collisions.

At the zonal level, Osama and Sayed (2016) developed a safety evaluation model using bike network indicators as predictors for cyclist safety. They focused on the city of Vancouver, British Columbia and used TAZs as the unit of analysis. Importantly, connectivity, directness and topography of the bike network were assessed as network indicators coupled with an exposure variable (bike kilometers travelled). The results showed that collisions involving cyclists are greatly influenced by bike network connectivity, directness and exposure. Bike network topography was found to have a negative association with cyclist collisions.

Ladron, de Guevara et al. (2004) developed planning level collision prediction models for the city of Tucson, Arizona. They used TAZs as the unit of analysis and calibrated prediction models using negative binomial regression. The research demonstrated that the use of such models can facilitate understanding of the factors that give rise to collisions by different levels of severity. They found that intersection density and the proportion of persons aged 17 years or younger expressed as a percentage of the total population could be used to predict the number of fatal collisions. Also, population density, number of people employed, and the length of arterials (major and minor) and collectors were found suitable for predicting injury and property damage only collisions.

Lovegrove and Sayed (2007) explored the predictive capabilities of collision prediction models developed in their previously reported study (Lovegrove and Sayed, 2006). The model was developed for the Greater Vancouver area and used negative binomial regression. The main purpose of the research was to evaluate the potential of using zonal level models in the identification of hotspots. The results demonstrated potential in the use of zonal level models for hotspot analysis. The results also showed that a zonal level approach to safety could be used to complement the traditional safety evaluation technique recommended by the Highway Safety Manual (2010). Other studies on the development of zonal level models include Pirdavani et al. (2012), Wei et al. (2011), and Lovegrove and Sayed (2006).

Takyi et al. (2018) recently used negative binomial regression in the development of crime and collision predictive models for the City of Regina. The research focused on analysis at the zonal level. Crime and collision prediction models were developed as a function of demographic, land use and socio-economic characteristics. The models developed were used in the identification of hotspots for enforcement prioritization.

The wide literature on the use of negative binomial regression validates its acceptance as the standard modelling technique for count data across various fields including criminology and collision analysis. This can be attributed to its ability to account for over-dispersion. However, the standard negative binomial model used for crime and collision modelling considers the parameter of independent variables fixed in explaining relationship across geographic surface. It does not consider spatial dependency in the data and the need to consider this spatial dependency while accounting for over-dispersion. This dissertation explores this opportunity for improvement in negative binomial regression models.

Poisson Log-normal Regression Models

The Poisson log-normal has been proposed as an alternative to the negative binomial model for count data modelling and has been applied in crime and collision studies. According to Fahrmeir et al. (2013), this type of model is particularly useful when predictor variables have an additive effect on the rate of occurrence of events (see Equation (2-2) for the form of the Poisson log-linear model). However, when this type of model accounts for over-dispersion, the expected mean $E(y|x)$ and the variances $Var(y|x)$ is obtained using Equations (2-11) and (2-12):

$$E(y|x) = \mu = exp(x_i^T \beta) \qquad \textit{(2-11)}$$

$$Var(y|x)=\theta\mu, \textit{ where } \theta = \frac{1}{\alpha} \textit{ is the over-dispersion.} \qquad \textit{(2-12)}$$

Brown and Oxford (2001) applied a Poisson log-normal model to modelling break and enter crimes in the city of Richmond, Virginia. In their study, the spatial distribution of these crimes was investigated. Brown and Oxford (2001) compared the performances of Poisson log-normal models developed with ordinary linear Poisson to cumulative logistic regressions models. The Poisson log-normal model was identified as suitable for modelling break and entry crimes in the City of Richmond, Virginia. It was also revealed that population variables also influence the number of break and entry crimes taking place at a location. For example, the proportions of persons aged 12 to 17 and 18 to 24 had a positive association with break and entry crimes. It was suggested that the Poisson distribution may not be suitable for modelling break and entry crimes when over-dispersion is considered.

Plassmann and Lott (2004) used a multivariate Poisson log-normal model to understand the relationship between gun ownership and violent crime (murder, rape and robberies). They based their analysis on data obtained from subscription for a gun magazine known as Guns and Ammo, a magazine popular among gun owners. The use of the data from this source was solely due to the difficulty in obtaining county level data on gun ownership. Magazine subscription records were related to the number of violent crimes at the county level. The multivariate Poisson log-normal regression showed a strong relationship between murder, rape and robberies and gun ownership.

To understand macro level factors that affect collisions, Aguero-Valverde (2013) used a Poisson log-normal model to study property damage only collisions at the canton level in Costa Rica. The model was found suitable for identifying factors that affect property damage only collisions. The multivariate Poisson model outperformed the univariate counterparts.

El-Basyouny and Sayed (2009) used a modelling approach similar to that of Aguero-Valverde (2013) to develop a collision prediction model for signalized intersections. The research focused on the use of multivariate Poisson log-normal modelling to determine hazardous areas (hotspots) in the City of Edmonton, Alberta. It was emphasized that accounting for extra variation in the Poisson model through the use of Poisson lognormal enhances the predictive performances of the models. Like Aguero-Valverde, El-Basyouny and Sayed reported that the multivariate Poisson log-normal model predicted collisions better than did the univariate versions.

El-Basyouny et al. (2014) used Bayesian multivariate Poisson log-normal regression in investigating the effect of time and weather on collisions in the City of Edmonton. They used five years of collision and daily weather data in their model development. It was reported that the multivariate Poisson log-normal provided a better fit to their collisions data. Also, it was found that temperature and snowfall were important variables influencing collisions rates. Collision rates were reduced with increased temperature but increased with snowfall intensity. A period of wet weather followed by a dry period was associated with stop sign violations, run off the road collisions, and following too close collisions. Variation in traffic volume during the days of the week was reported to play a significant role in understanding the effect of weather on traffic collisions.

Wang and Kockelman (2013) adapted the Poisson log-normal to the conditional autoregressive version and used it in modelling pedestrian collisions at the census tract level in Austin, Texas. A pedestrian collisions prediction model was calibrated as a function of land use, demographic and network characteristics. Conditional autoregressive modelling allows heterogeneity and correlation in the response variable across space to be accounted for. The results suggested that mixed land use increases the risk of pedestrian collisions for some severity levels. Also, a high number of access points within a zone was associated with an increase in the number of conflict movements between pedestrian and vehicles. However, the provision of side walks enhanced pedestrian safety.

Similar applications of multivariate Poisson log-normal regression have been reported by Wang et al. (2017) in modelling collisions by severity and types on rural highway and also by Zhang et al. (2015).

Interestingly, most research that has used Poisson log-normal regression techniques has compared the univariate and multivariate versions of the models, especially in the case of transportation safety research. In all cases, the conclusions favoured multivariate Poisson log-normal regression modelling and accounting for the over-dispersion characteristics of count data.

Zero Inflated Regression Models

Zero inflated regression models may be used when the data used includes records of zero. In some cases, the presence of zeros can be attributed to differences in observation conditions. It may imply low likelihood of an event occurring. Such observations are usually characterized by two states. The first state is the normal count where the likelihood of incidents occurring is non-zero. The second state refers to conditions where the likelihood is zero (Washington et al., 2003). When a data collection sequence is characterized as a normal and zero state, zero inflated Poisson or negative binomial regression is usually used to account for the dual states.

The zero inflated Poisson regression model follows the formulation of the traditional Poisson model except for accounting for two phases. If the zero and non-zero phase occurred with probability of π and $1-\pi$, the probability distribution of zero inflated Poisson model can be given as Equation (2-13):

$$P(y_i = j) = \begin{cases} \pi_i + (1-\pi_i)exp(-\mu_i) & if\ j = 0 \\ (1-\pi_i)\frac{exp(-\mu_i)\mu^y}{y_i!} & if\ j > 0 \end{cases} \qquad \textit{(2-13)}$$

where π_i is the link function.

Similarly, the zero inflated negative binomial probability distribution function is expressed as 2-14:

$$P(y_i = j) = \begin{cases} \pi_i + (1 - \pi_i) f(y_i = 0) & if\ j = 0 \\ (1 - \pi_i)\ f(y_i) & if\ j > 0 \end{cases} \qquad \textit{(2-14)}$$

Depending on the value y_i, the probability density function of the zero inflated negative binomial model is expressed as Equation (2-15):

$$f(y_i) = P(y_i|\mu_{i,}\alpha) = \frac{\Gamma\left(\frac{1}{\alpha}+y_i\right)}{\Gamma(y_i+1)\Gamma\left(\frac{1}{\alpha}\right)} \left(\frac{1}{1+\alpha\mu_i}\right)^{\frac{1}{\alpha}} \left(\frac{\alpha\mu_i}{1+\alpha\mu_i}\right)^{y_i} \qquad \textit{(2-15)}$$

Both the zero inflated Poisson and negative binomial model has been found useful in modelling crime and collision count data.

In a study designed to improve understanding of the relationship between police shootings, race and violent crime, Klinger et al. (2015) used zero inflated Poisson regression. They focused their research on the City of St Louis, Missouri and investigated whether police shootings correlated with areas experiencing high level of violent crimes and the racial structure of the environment. The study focused on records from 2003 to 2012. The results suggested that race and socio-economic characteristics at the neighbourhood level did not positively influence police shootings. Police shootings were less prevalent in areas with a high record of violent crime in St Louis. Klinger et al. (2015) recommended that the findings should be validated in other areas.

Disorderliness refers to a loss of societal control that affects individuals and neighbourhoods' characteristics as measured by structural characteristics of a neighbourhood such as the poverty level, economic inequality and public incivilities (e.g., drinking, illegal drug use and vandalism) (Simpson and Raudenbush, 2001; Wallace and Scott, 2017). The relationship between disorderliness and crimes has been investigated using zero inflated negative binomial regression.

Yang (2010) investigated the spatial association between disorderliness and violent crime in Seattle, Washington and Washington D.C. Yang reported that social disorderliness is not random in space, but forms clusters which could be deemed hotspots. More interestingly, using a zero inflated Poisson regression model, disorderliness in society was highlighted to have a 30% probability of predicting violent crimes.

Gover et al. (2008) used zero inflated Poisson regression to understand the lawbreaking behaviours (escape, attempted escape, fighting, etc.) of convicted male and female persons while in correctional facilities. This research was done to support theories of how well inmates adjust to the prison environment. Gover et al. used 247 official records and self-reporting data of persons incarcerated. In the study, the record of inmate misconduct behaviour was considered a response variable. The researchers treated characteristics of the inmates (age, race, education, length of imprisonment, and type of offense (violent or non-violent) and type of correctional facility (minimum or maximum security) as predictors of the behaviours of inmates. Findings from the research indicated that the predictors of male and female behaviours of persons serving prison times are different. Previous sentencing and jail time affected the behaviour of both male and female inmates, but in different ways. Previous sentencing was positively associated with law-breaking behaviour of male inmates. The length of the prison sentence significantly influenced the unlawful behaviour of both male and female inmates. The security level of the prison also impacted the unlawful behaviours of the prisoners serving jail time. Most importantly, the offence which led to sentencing was key to inmates' criminal behaviour. Gover et al. recommended gender specific programs that focus on helping inmates to improve their behaviours while in prison.

Zero inflated Poisson regression methods have found few applications in traffic collision prediction. Lee et al. (2002) emphasized that while zero inflated Poisson regression models have been introduced for handling zeros in count data, the models have not been extensively applied in collision studies. Lee et al. (2002) evaluated factors predicting young driver motor vehicle collisions using zero inflated Poisson models as a function of the demographic, driving experience and behavioural characteristics of young drivers. It was revealed that the likelihood of young driver being involving in one or more collisions in the first 12 months of driving was dependent on driving experience gained before obtaining their license. Comparison of the model developed with negative binomial regression was reported to produce the same result. Interestingly, dispersion was not thoroughly examined in the research as the models developed were more explanatory than predictive. It was recommended that different modelling options should always be considered when dealing with data containing excess zero. This is to ensure that the best model that properly represents the data is selected.

Carvalho and Lavor (2008) used a zero inflated negative binomial model to investigate the effects of socio-economic characteristics of individuals and areas on repeated victimization of property by criminals in Brazil. They focused on the influence of income inequality on crime. Cavalho and Lavor found that inequality of income played a significant role in repeated property victimization by criminals in an area. This was consistent with what has been reported in previous literature.

Akyus and Armstrong (2011) used zero inflated negative binomial regression to understand factors that influence terror crimes in 81 provinces of Turkey. The predictor variables were poverty (an indicator for socio-economic characteristics), residential mobility (the extent to which people in certain areas migrate to other places), and the ethnic complexity/diversity of the provinces considered. Diversity within the ethnic populace of provinces was a major predictor of terror crimes when poverty and rates of residential migration were treated as control variables.

Zero inflated negative binomial regression has been applied in collision studies at the micro level. Li et al. (2008) used this type of model to understand the factors that influence and predict street car racer collisions. Focusing on Utah in the United States, models were developed as a function of the age, gender, environment where the collisions were recorded, and the number of previous speed citations received by the drivers. The results indicated that drivers who had records of prior speeding citation and poor driving history were likely to be involved in collisions. This risk was higher in drivers with previous speeding citations. Those without citations were considered to drive safely and are less likely to be involved in collisions.

In all the global models described, spatial dependency was not taken into consideration. Rather, a fixed parameter was estimated from the model and used to describe relationships across geographic space despite the problem that spatial dependency can cast doubt on the assumption of independence of models such as Poisson, Poisson log-normal, negative binomial and the zero inflated regression.

Accounting for spatial dependency requires the use of geographically weighted regression models.

2.4.2 Geographically Weighted Regression Models for Count Data

Location is one of the important factors that plays a role in the frequency of crime and collisions due to the variation in sociological, economic and demographic variables that influence crime and collision occurrence. Crimes and collisions are not homogenous over geographic surfaces. They are heterogeneous in nature and vary with sociodemographic and land use factors.

Recognition of spatial dependency has been attributed to Tobler’s law that relationships exist across geographic space. Recognition of the issue has cast doubt on the accuracy of fixed parameter models such as the Poisson and negative binomial models. This has led to the development of spatial models that take into consideration the influence of nearby areas.

The increasing recognition of the spatial dependency and heterogeneity that characterizes data such as crime and collision data has led to much research in the last decades, and the geographically weighted regression method has been proposed (Loo and Anderson, 2015). Geographically weighted regression is based on the idea that parameters can be estimated for different location within a study area provided that a set of dependent and independent variables is given (Charlton and Fotheringham, 2009; Huang et al. 2010). It examines relationships that exist across geographically referenced observations and the factors causing them. Geographically weighted regression is based on the idea that factors causing crime and collisions are not stationary but vary geographically.

The first version of this type of model was proposed by Brunsdon et al. (1996). It assumes a Gaussian distribution for the response variable. The functional expression of the geographically weighted regression model is given as Equation (2-16):

$$y_i = \beta_0(u_j, v_j) + \sum_{i=1}^{k} \beta_i(u_j, v_j) x_{ij} + \varepsilon_j \qquad \textit{(2-16)}$$

where:

u_j, v_j , represents the coordinate at which observation i is taken;

y_i and x_{ij} are the dependent and the independent variables respectively;

$\beta_0(u_j, v_j)$ and $\beta_i(u_j, v_j)$ represent the intercept and the estimated locally varying coefficient; and

k, and ε_j, are the number of independent variables and the error term for observation location j. Equation (2-16) can be written as (2-17) when exposure t_j, which represent the interval at which the event is studied, is taken into consideration.

$$y_i \sim t_j exp\left(\sum_{i=1}^{k} \beta_i(u_j, v_j) x_{ij}\right) \tag{2-17}$$

The concept of geographically weighted regression is established based on non-parametric smoothing and curve fitting techniques in which local regression parameters are determined from a subset of data around estimation points. It uses data within a neighbouring location in a geographic space for model calibration to explore spatially varying relationships (Wheeler and Paez, 2010). Most importantly, geographically weighted regression assumes that observations closer to each other influence parameter estimates of a data point in a geographic space, thus incorporating the first law of geography – everything is related to everything, but closer things are more related (Thapa and Estoque, 2012)

Geographically weighted regression as a non-parametric technique that uses subsampling of observed data across space for statistical analysis and has become important in the study of spatially referenced data such as crime and collision data. However, the concept of subsampling statistical data, though innovative in spatial statistics, is not new to classical statistics. Its application in the field of spatial statistics may be attributed to Brunsdon et al.'s (1996) research that applied the concept of variable geographic space.

The original reference to the concept used by Brunsdon et al. (1996) could be linked to a smoothing technique for histograms. The idea of using distance weighting also existed in interpolating algorithms, but the advancement of weighted concepts in geography paved the way for a multivariate local spatial data approach. The advantages of understanding local relationships and patterns have made the geographically weighted regression more popular (Paez et al., 2011). However, it can still be argued to be an extension of locally weighted regression like quantile regression which explores the variation in relationships among statistical data. Similarly, geographically weighted regression is related to conditionally parametric regression which is an extension of locally weighted regression. The difference is that the parameters (coefficients) of geographically weighted model are non-parametric estimates from longitudes and latitude, i.e. the distance between observations and target points of a locally weighted regression (McMillen, 2012).

The need to accommodate spatial dependency and heterogeneity has led to increased use of geographically weighted regression and has become a method for modelling spatial non-stationarity (Whigham and Hay, 2007). It has found wide application in various fields of study. Areas in which geographically weighted regression has been used include urban planning (Huang et al, 2010; Du and Molley, 2012; Kyratso, and Yiorgos, 2004), demographic studies (Bajat et al, 2011), health sciences (Comber et al., 2011; Lin and Wen, 2011), ecological biology (Windle et al, 2009), social sciences, crime studies (Cahill and Mulligan, 2007), and transportation studies (Hadayeghi et al., 2010; Zhang et al., 2015; Zheng et al., 2011). However, the geographically weighted regression proposed by Brunsdon et al. (1996) in not suitable for non-negative count data such as crime and collision data due to the Gaussian distribution placed on the response.

In the last decade, the geographically weighted regression method has been advanced to accommodate a Poisson and negative binomial distribution suitable for count data. These models are known as geographically weighted Poisson regression (GWPR) and geographically weighted negative binomial regression (GWNBR).

Geographically Weighted Poisson Regression (GWPR)

Geographically weighted regression has been extended from the originally proposed version by Brunsdon et al. (1996) to allow for a Poisson distribution on the response variable, thus making it suitable for count data that are not over-dispersed. This model is known as the GWPR and was proposed by Nakaya et al. (2005). It uses a conditional spatial kernel weighting function for the estimation of the parameter variation of the Poisson regression method (see Equation 2-18 for functional form of the GWPR):

$$y_i \sim Poisson\left[t_j exp\left(\sum_{i=1}^{k} \beta_i\left(u_j, v_j\right) x_{ij}\right)\right] \qquad (2\text{-}18)$$

The parameters in Equation (2-18) are the same as those of (2-16) and (2-17) except that the model in Equation 18 follows a Poisson probability density distribution.

In a recent application of GWPR in the development of a collisions prediction model by Pirdvanni et al. (2014a), it was established that it outperformed the traditional negative binomial regression known to be the state of art modelling in road safety. This was attributed to the advantage of the GWPR model, i.e., capturing different patterns (spatial dependency/heterogeneity) in the relationship between the aggregated collisions and the socio-demographic predictors.

Similarly, Li et al. (2013) used the GWPR to model the relationship between collisions and their spatial correlates using California data as a case study. Results consistent with the study by Pirdvanni et al (2014a) were obtained and GWPR model was found useful for capturing the spatial relationship between collisions and reducing spatial correlation in the residual obtained from the models developed.

The GWPR model has also been used to explore the impact of teleworking on the frequency of collisions in the Flanders region of Belgium. The traffic safety benefit of teleworking was highlighted using this approach. As expected, teleworking reduced the total vehicle kilometers travelled which resulted in a reduction in vehicular collisions (Pirdvanni et al 2014b).

Most applications of the GWPR model are found in transportation studies. Most researches that have used geographically weighted regression in crime studies have still used the type proposed by Brunsdon et al. (1996).

Interestingly, most researchers that have used the GWPR model have compared the performances of the model developed with negative binomial regression and have reported the advantages of improved performance. However, the problem of over-dispersion that characterizes count data is downplayed in GWPR model.

While accounting for spatial dependency could enhance the accuracy of models, over-dispersion in count data is not considered by the GWPR and could constitute a problem.

Geographically Weighted Negative Binomial Regression (GWNBR)

Another extension of geographically weighted regression is the GWNBR. This model is based on the idea that the over-dispersion that characterizes spatially referenced count data such as crime and collision data should be taken into consideration. The GWNBR was proposed by Da Silva and Rodriguez (2014). The concept of this model is similar to that of negative binomial regression. However, rather than using a fixed parameter to describe relationships across space, the parameters vary locally. This is similar to the GWPR except that over-dispersion is accounted for. The functional form of the GWNBR model is given as Equation (2-19):

$$y_i \sim NB\left[t_j exp\left(\sum_{i=1}^{k} \beta_i\left(u_j, v_j\right) x_{ij}\right), \alpha \right] \tag{2-19}$$

The parameters in Equation (2-19) are similar to those in (2-18) except that α represents the over-dispersion parameter.

Geographically weighted Poisson regression and the traditional version, which assumes a Gaussian distribution for the response, was used extensively in crime and collision studies, but over-dispersion had not been thoroughly treated until Da Silva and Rodriguez (2014) proposed the GWNBR model. The proposed method has not yet been comprehensively evaluated to determine the improved performance offered compared with the GWPR model. This is a research gap that needs to be investigated.

2.5 Bandwidth Influence on Geographically Weighted Regression

The bandwidth is important to effective calibration of geographically weighted regression methods (Zhen et al., 2013). However, different types of bandwidth exist in the study of local relationships across geographic space. The most commonly used bandwidth types are the fixed Gaussian bandwidth and the adaptive bi-square bandwidth. The choice selected affects the performance of geographically weighted models (Jacquez, 2010). This is because the bandwidth selected affects the extent to which the coefficient of the models varies (Bitter et al., 2007; Blainey and Mulley, 2013; Du and Mulley, 2012; Timofeeva and Tesselkina, 2016).

The optimal bandwidth represents the threshold at which neighbouring units have an influence on locally calibrated parameters (Shi et al., 2006; Su et al., 2017). Thus, the number of nearest neighbours selected affects the distribution of the coefficients. A larger bandwidth produces a global regression model that uses a fixed coefficient to describe a relationship across geographic area. On the other hand, a smaller bandwidth captures local relationships much better (Su et al., 2017).

Some researchers have evaluated the performance of geographically weighted regression considering different bandwidth types. Bidanset and Lombard (2014a) explored used real estate data in an evaluation of the uniformity of the selling prices of houses in Norfolk, Virginia. They examined the sensitivity of locally varying regression models to bandwidth specifications. Most importantly, they focused on geographically weighted regression that assumes a Gaussian distribution for the response variables. Bidanset and Lombard compared the performance of the two most used bandwidth methods (fixed and adaptive). It was reported that geographically weighted regression with fixed bandwidth produced a better fit in modelling the real estate prices. For their data, the model with adaptive bi-square bandwidth produced lower performance results when measured by the Akaike information criterion. Caution was recommended in the use of predetermined bandwidths and researchers should endeavor to compare bandwidth types. This is to facilitate the best performances and a better justification of the model selected. Similar findings and recommendations were reported by Bidanset and Lombard (2014b).

In contrast, Chasco et al. (2007) studied household disposable income using geographically weighted regression models in Spanish provinces. They compared the fixed and the adaptive version of the Gaussian and bi-square bandwidth function. The results obtained were contrary to what was reported by Bidanset and Lombard (2014). Chasco et al. found adaptive bi-square kernel bandwidth function to produce the best fitting geographically weighted regression. Interestingly, both studies used a method of cross-validation in their optimal bandwidth selection.

Bitter et al. (2007) used adaptive Gaussian bandwidth geographically weighted regression in the examination of spatial variation in house prices in Tucson. Although they arbitrarily chose a type of bandwidth for their analysis, geographically weighted regression was found to produce a better fit to their housing price data than the global model. The spatial structure of urban settings has also been modelled by geographically weighted regression using preselected fixed bi-square bandwidth (Noresah and Ruslan, 2009).

The variation reported in geographically weighted model performances by bandwidth choice suggests that different bandwidth types may perform differently with different data and different methods of optimal bandwidth selection.

While the two most commonly used common types of bandwidths are fixed Gaussian and the adaptive bi-square, most researchers have preferred the adaptive bi-square bandwidth over the fixed Gaussian bandwidth. Research that has used adaptive bandwidth in calibrating geographically weighted regression has followed Fotheringham et al.'s (2002) approach. Fotheringham et al. stated that while the geographically weighted regression method could be influenced by the bandwidth choice, adaptive bandwidth performs better when the data include sparse and dense data distributions. This assertion has been used across various reported studies in transportation engineering without comparing the effect of different bandwidths.

It is of interest that the performance of these bandwidths may vary with different distributional assumptions, and that the data structure may play a key role. The idea that both methods should be evaluated should always be considered. This is to allow accurate choice of bandwidth for geographically weighted regression analysis.

The approached used to determine the optimal bandwidth in the calibration of geographically weighted regression methods affects the optimal bandwidth selected. The approaches used include cross-validation and the Akaike information criterion (Li et al., 2010). In Farber and Paez's (2007) study of methods used to determine the optimal bandwidth for calibration of geographically weighted models, the researchers found that the majority of the 64 papers examined used cross-validation. Precisely, 35 of 64 studies used the cross-validation method, 13 used Akaike information criterion, 9 used a predefined method, and 7 did not mention the method used (the method was possibly chosen arbitrarily). Farber and Paez suggested that cross- validation is the preferred approach for optimal bandwidth determination. This is supported in other research into the most suitable approach (Cho et al., 2010; Griffith, 2008; Timofeev and Tesselkina, 2016; Wang et al., 2008). Cho et al. (2010), for example, used agricultural data and compared the performance of two methods (cross- validation, and smallest spatial error Lagrange Multiplier) for optimal bandwidth selection. It was reported that the geographically weighted regression from the cross-validation method produced a better fit than did the smallest spatial error Lagrange Multiplier.

2.6 Test for Spatial Dependency

Application of spatial models such as the geographically weighted regression model in crime and collision studies requires the presence of spatial dependence in data to be established. Various tests have been developed to determine spatial dependency, and the most commonly used are the Moran I and Geary C tests.

According to Radil (2016), both the Moran I and Geary C tests can be applied to determine the presence of spatial dependence in continuous interval or ratio-level measurements. These methods were initially global measurements of spatial dependency. This is because they determined the divergence of variables from randomness across geographic space using a single statistic. Interestingly, the two methods use different approaches for estimating dependence across variable space.

The Moran I test determines the variation in spatial randomness using an approach similar to Pearson correlation, i.e., the cross product of the coefficients. It can vary from +1 which indicates the presence of perfect positive spatial dependence to -1 which implies the presence of negative spatial dependence.

The Geary's C test uses the square of the deviation in the estimation of spatial dependency. However, in this case, a value less than 1 implies positive spatial dependence, and values greater than 1 indicate negative spatial dependence.

There are local versions of these tests for spatial dependency. The local Moran I and local Geary's C tests for spatial dependence evaluate the existence of local clusters within the study regions. The Local Moran I was proposed by Anselin (1995) and the local Geary's C (Gi*) statistics introduced by Getis and Ord (1992).

These local tests of spatial dependence compare values occurring within proximities. In all cases, they evaluate the null hypothesis of absence of spatial randomness against the presence of spatial randomness otherwise known as spatial dependence. The more commonly used local measure for spatial dependence is the local Moran I statistic (Ratcliffe, 2010). This is because it is flexible and can be adapted to different situations. For example, it could be applied to a situation where information is presented in areal units or as a point pattern. Its advantage of accounting for underlying variations of process makes it the preferred choice for researchers (Bernasco and Elffers, 2010). Asides, Moran I is computationally robust, and the interpretation of the results is intuitive i.e. positive and negative Moran I indicate presence of positive and negative spatial dependency.

2.7 Adjusting for Regression-to-the-Mean

Different methods exist for correcting RTM bias. The commonly used approaches include the Continuous Risk Profiling (CRP) technique and the Empirical Bayes method (Lee et al., 2016). The approach selected for correcting RTM bias depends on expertise and the choice preferred for a research project. In some cases, selection is based on the standard relied on in the identification of hotspots.

Continuous Risk Profiling (CRP) is a technique proposed by Chung and Ragland (2007) to address the problem of false positives, i.e. incorrect identification of hotspot locations for safety improvement (Kim et al., 2014). It is used to correct for the RTM bias that characterizes before-and-after observational studies. The method is based on observed collisions and has been used to identify hotspot locations needing safety improvement. This approach generates a risk propensity (also known as a profile) for road segments within a study area based on a weighted moving average of the collisions that occurred on each segment. The profile obtained is then compared with the baseline threshold and areas that exceed this threshold are deemed hotspots that require immediate attention (Medury and Grembek, 2016). CRP adjusts to the true risk potential of a site by estimating the density of collisions. It focuses mainly on changes in collision risk potential at sites, thus providing an opportunity for effective evaluation of hotspot countermeasures designed to improve roadway safety (Chung and Ragland, 2009).

According to Kim et al. (2014), CRP methods use three main steps when identifying hotspots. Firstly, random variation in collisions data is removed using the technique known as the weighted moving average. Secondly, the collision data and prediction models are used to determine the boundary of the sites. Thirdly, collisions within the defined boundary are then used to select sites that will receive countermeasures. CRP has been used extensively in the identification of collision hotspots for countermeasure provision.

Oh et al. (2009) used a CRP method to determine characteristic of hotspot locations that experienced high collision frequency during wet roadway pavement conditions. The method was used solely to minimize exaggeration or perhaps misidentifications of collision hotspots while determining factors contributing to the collisions. The false positives, i.e. the misclassifications obtained from the CRP method, were compared with another method known as the sliding window technique. Interestingly, the CRP method was reported to outperform the sliding window technique. A 30% reduction of false positives was recorded. This method was found to account for a shift in hotspot locations during dry and wet pavement conditions. This shift in hot spot location can be attributed to the stochastic variation in collision incidences.

Similarly, Chung et al. (2011), used the CRP technique in the identification of sites for improvements on freeways. Chung et al. (2011) proposed a method that combines the CRP technique with a cumulative sum algorithm to evaluate safety deterioration on freeway segments. Safety deterioration was measured as changes in the observed collision rates. Essentially, the cumulative sum algorithm was used to detect sudden variation in the collision records. Chung et al. found that by using a growth factor determined by the cumulative sum algorithm and observed changes measured using CRP, locations where safety is gradually deteriorating could be determined and countermeasures provided. The research by Chung et al. showed that the CRP method can be proactively used to identify hotspots. Interestingly, the CRP method has been used extensively to analyze roadway segments.

The Empirical Bayes (EB) method is the most popular method for reducing RTM bias in hotspot identification. The EB method uses information from a group of entities (the reference group) combined with historical collision data to determine the safety of a road segment or intersection. The approach uses detailed mean and variance information for a similar group of entities (Hauer, 1992). By incorporating mean and variance information from the reference group, the EB method can control random variation in the number of collisions recorded at a site. This method has been reported to handle RTM bias well (Yang and Loo, 2016, Hauer et al., 2002) and has identified hotspots in a stable manner, i.e., it increases precision (Cheng and Washington, 2008; Hauer et al., 2002; Montella, 2010).

However, the initial EB method required a large reference group and the group selected could be somewhat arbitrary. The reference groups can be difficult to match to the site being investigated. Due to this limitation, a multivariate regression approach that can estimate the expected mean and variance (dispersion) was proposed by Hauer (1992). Application of regression methods to estimate the mean and variance of reference group makes EB a flexible way to estimate safety and identify hotspots. This flexibility is not limited to the safety analysis of road segments and could also be found applicable in zonal level (macro level) safety analysis.

The EB method has been widely applied in safety analysis in transportation engineering. It has been recognized as the statistically preferred approach for before-and-after evaluation of safety countermeasures (Persaud and Lyon, 2007). Other researchers have used EB to focus on accurate identification of hotspots.

At the segment level (micro level analysis), Persaud et al. (2004) evaluated the safety effects of lane separation on two-lane rural roadway using data obtained from California, Colorado, Delaware, Maryland, Minnesota, Oregon and Washington. In total, 98 sites along 210 miles (about 338 kilometers) of road segment was considered. The research was motivated by the need to propose safety measures to mitigate collisions occurring on rural highways with no physical separations (e.g., no median or barrier between lanes). Collisions between vehicles travelling in opposing directions account for the majority of fatalities on rural highways. Persaud et al. evaluated the safety implications of installing rumble stripes in the middle of rural highways to warn vehicles that crossed the centreline, often because of distraction or fatigue. Using an EB before-and-after study, the effect of rumble strips was assessed on selected hotspots. The results indicated that the rumble strips brought about a significant reduction in injury related collisions. Using the empirical Bayes approach accounted for RTM and meant that the safety effect was not exaggerated.

Shaheed et al. (2015) used the EB technique for the identification of collision hotspots in Iowa during winter using collisions data from 2008 to 2009 and from 2011 to 2012. Winter collisions were extracted. The analysis incorporated weather-related factors such as weather conditions, surface conditions and visibility at the time of collision as predictors in the model development. Negative binomial regression was used in the development of the collision prediction model. The EB method identified high risk collision segments accurately and was able to overcome the challenges associated with traditional crash analysis techniques.

Cafiso and Silvestro (2011) used a similar technique to identify black spots/hotspots on two-lane local rural roads where only sparse data were available. Due to the criticality of the predictors (e.g., traffic volume and segment length) and the collision observation period of the data collected, Monte Carlo simulation was used to produce theoretical collision data similar to the empirical and a priori data used to define the hotspots. The results obtained from the EB method were compared with observed the collision frequencies, the crash rate and the potential for safety improvement. The results showed that the EB estimates were the best approach for identifying hotspots on rural highways characterized by sparse data. The simulation approach was recommended for constructing collision data similar to empirical data when evaluation of hotspots on two lane local rural road ways is to be carried out.

The EB method has also been applied in the identification of animal vehicle collision hotspots. Gkritza et al. (2013) used this technique to assess safety improvement needs on 150 highway sections in Iowa using records of deer-vehicle collisions. A SPF was developed using a negative binomial model that regressed deer vehicle collisions as a function of roadway characteristics and environmental factors. Analysis using the EB method was able to identify a list of the top 25 roadway segments in need of safety improvements. Most were located on roadways with high speed, right shoulder areas and adjacent grassland. The analysis facilitated the provision of deer vehicle collision countermeasures. The research showed that the application of the EB method is not limited to the identification of vehicle to vehicle collision hotspots only.

Although the EB method has been identified as a technique that enhances hotspot identification, and the EB method has received much attention at the microscopic level, its application in macroscopic modelling is still lacking. Most literature on the application of the EB technique is at the microscopic level and involves developing a SPF and incorporating the SPF estimate into the EB technique. Application of methods for correcting the biases associated with RTM in macroscopic collision studies is still a grey area.

Macro-level prediction models can also be used in the EB identification of hotzones. The use of the models will allow for spatial factors such as demographic and land use characteristics to be accounted for in crime and collision prediction and incorporating the EB method into such analyses could lead to improved predictive performance. The approach could also lead to better identification of hotzones for effective deployment of enforcement resources.

Comparison of Methods for Adjusting for Regression-to-the-Mean

Various researchers have evaluated the performance of CRP as a method for overcoming the variations that surround collision frequency. Some researchers have evaluated the advantages of the CRP method over similar approaches such as the sliding window methods that focus on road segments (Grembek et al., 2012; Lee et al., 2013; Oh et al., 2009).

The CRP method is not without limitations. The restrictions of this approach to safety evaluation and hotspot identification at segment level are a strong limitation. Chung et al. (2009) explained that the CRP method is most suitable for the determination of hotspots on continuous road way sections such as freeways. The idea that CRP is suitable for continuous sections of roadway was also highlighted in the research by Lee et al. (2016) in their study comparing the performance CRP with the EB method. Lee et al. used 110 miles of freeway in California. The results were consistent with the claims that CRP is the most suitable method for continuous sections of roadways. It was emphasized that CRP outperformed the EB method in identification of hotspot. They concluded that the performance of the EB method could be limited by a SPF that is biased. The CRP technique in the evaluation of safety is limited to continuous sections. Intersection studies remain limited.

Srinivasan et al. (2011) argued that the proposed CRP technique uses only observed collisions in the determination of hotspots for countermeasure provision and therefore does not correct for RTM bias.

The alternative approach to the CRP method is the EB technique. The EB technique has been compared with similar methods such as simple ranking methods and its advantages emphasized. Cheng and Washington et al. (2005) compared the predictive performance of a model developed using the EB technique with a confidence interval approach to ranking sites. Using experimental simulated data, Cheng and Washington et al. argued that the EB method is superior to the similar approaches (i.e., superior to simple ranking and the confidence interval method).

A similar comparison was carried out by Montella (2010) and revealed that the EB method outperformed other hotspot identification methods. The EB method was found to be consistent and reliable in the identification of hotspots. Emphasis was placed on the consistency of this finding with previous research.

Elvik (2008) examined the predictive performance of the EB method for road safety using estimates derived from empirical distributions of collisions at sites, and predictions obtained from a negative binomial model fitted to the empirical distributions. The estimates were combined in the EB analysis. The analysis showed that incorporating the EB method into regression analysis enhances prediction of collisions, i.e. it improves precision. This again highlights the advantages of the EB method. The EB method has thus become the state-of-the-art approach to collision hotspot identification.

Apart from the precision offered by the EB method, it is more flexible than other approaches such as the CRP. The EB method is not restricted to road segments. It can be applied to various components of the transportation network including segments and intersections. Its advantages are not limited to microscopic level analysis¸ but can be applicable in macro-level analysis to correct for RTM bias and enhance precision. The EB method has numerous other advantages. It allows the use of predictive models known as SPFs in establishing the relationship between collisions and their associated factors. Furthermore, it allows information from observed collision frequency to be used to enhance the identification of hotspots, thus accounting for random variation in collisions frequency known as RTM. The EB method is the technique recommended by the Highway Safety Manual (2010) and most used by researchers. Thus, the EB method will be used for correcting for RTM in this research.

In summary, this chapter has provided a background into the concepts involved in a DDACTS analysis. It identified the spatial units of analysis at which hotspots could be determined, i.e., both micro-level and macro-level, as well as the applications, but emphasis was placed on the use of macro-level analysis. The chapter highlighted TAZs as the spatial unit of analysis for macro-level analysis preferred by transportation professionals and it was the unit considered in this research. The chapter also discussed the problem of spatial dependency and heterogeneity in spatially referenced data such as crime and collision incidences, the regression methods for count data modelling and the regression-to-the-mean problem. The need for spatial regression methods such as the geographically weighted regression method to address the problem associated with spatially count crimes and collision was emphasised. This research focused extensively on the geographically weighted regression method Also, the influence of bandwidth on the geographically weighted regression models and method for testing for spatial dependency was highlighted. Lastly, the methods for correcting for RTM bias were thoroughly compared. However, the advantages of the EB method highlighted made it a preferred approach used in this research.

References

Abdel-Aty M, Lee J., Siddiqui C, and Choi K., (2013). Geographic unit-based analysis in the context of transportation safety planning. Transportation Research Part A. Vol. 49, pp. 62-75

Abdel-Aty, M. A., and Radwan, A. E. (2000). Modeling traffic accident occurrence and involvement. Accident Analysis and Prevention, 32(5), 633-642.

Aguero-Valverde, J. (2013). Multivariate spatial models of excess crash frequency at area level: case of Costa Rica. Accident Analysis and Prevention, 59, 365-373.

Akyuz, K., and Armstrong, T. (2011). Understanding the sociostructural correlates of terrorism in Turkey. International Criminal Justice Review, 21(2), 134-155.

Alkahtani, K. F., Abdel-Aty, M., and Lee, J. (2018). A zonal level safety investigation of pedestrian crashes in Riyadh, Saudi Arabia. International Journal of Sustainable Transportation, 1-13.

Anselin, L. (1995). Local indicators of spatial association—LISA. Geographical analysis, 27(2), 93-115.

Anselin, L. (2003). Spatial externalities, spatial multipliers, and spatial econometrics. International regional science review, 26(2), 153-166.

Autey, J. (2012). Before and after traffic safety evaluations using computer vision techniques (Doctoral dissertation, University of British Columbia).

Bajat, B., Krunić, N., Kilibarda, M., and Samardžić-Petrović, M. (2011). Spatial modelling of population concentration using geographically weighted regression method. Journal of the Geographical Institute Jovan Cvijic, SASA, 61(3), 151-167

Barnett, A. G., Van Der Pols, J. C., and Dobson, A. J. (2004). Regression to the mean: what it is and how to deal with it. International journal of epidemiology, 34(1), 215-220.

Basile, R., Durbán, M., Mínguez, R., Montero, J. M., and Mur, J. (2014). Modeling regional economic dynamics: Spatial dependence, spatial heterogeneity and nonlinearities. Journal of Economic Dynamics and Control, 48, 229-245.

Beland, L. P., and Brent, D. A. (2018). Traffic and crime. Journal of Public Economics, 160, 96-116.

Bernasco, W. (2010). Modeling micro-level crime location choice: Application of the discrete choice framework to crime at places. Journal of Quantitative Criminology, 26(1), 113-138.

Bernasco, W., and Elffers, H. (2010). Statistical analysis of spatial crime data. In Handbook of quantitative criminology (pp. 699-724). Springer, New York, NY.

Bidanset, P. E., and Lombard, J. R. (2014a). The effect of kernel and bandwidth specification in geographically weighted regression models on the accuracy and uniformity of mass real estate appraisal. Journal of Property Tax Assessment & Administration, 10(3).

Bidanset, P. E., and Lombard, J. R. (2014b). Evaluating spatial model accuracy in mass real estate appraisal: A comparison of geographically weighted regression and the spatial lag model. Cityscape, 16(3), 169-182.

Bitter, C., Mulligan, G. F., and Dall'erba, S. (2007). Incorporating spatial variation in housing attribute prices: a comparison of geographically weighted regression and the spatial expansion method. Journal of Geographical Systems, 9(1), 7-27.

Blainey, S., and Mulley, C. (2013, October). Using geographically weighted regression to forecast rail demand in the Sydney region. In Australasian Transport Research Forum.

Boivin, R., and Felson, M. (2018). Crimes by visitors versus crimes by residents: The influence of visitor inflows. Journal of Quantitative Criminology, 34(2), 465-480.

Britt, C. L., Rocque, M., and Zimmerman, G. M. (2018). The analysis of bounded count data in criminology. Journal of Quantitative Criminology, 34(2), 591-607.

Brown, D. E., and Oxford, R. B. (2001). Data mining time series with applications to crime analysis. In Systems, Man, and Cybernetics, 2001 IEEE International Conference on (Vol. 3, pp. 1453-1458). IEEE.

Brunsdon, C., Fotheringham, A.S., Charlton M.E., (1996). Geographically weighted regression: a method for exploring spatial non-stationary. Geographical Analysis, 28(4), 281-298.

Brunsdon, C., Fotheringham, S., and Charlton, M. (1998). Geographically weighted regression. Journal of the Royal Statistical Society: Series D (The Statistician), 47(3), 431-443.

Burch, J. H., and Geraci, M. N. (2009). Data-driven approaches to crime and traffic safety. The Police Chief, 76(8), 18-23.

Burrell, G., Thoemmes, F., and MacKinnon, D. (2010). Visual displays of regression toward the Mean using SAS SGplot. In SAS Global Forum (Vol. 2010).

Cafiso, S., and Di Silvestro, G. (2011). Performance of safety indicators in identification of black spots on two-lane rural roads. Transportation Research Record: Journal of the Transportation Research Board, (2237), 78-87.

Cahill, M and Mulligan G (2007). Using geographically weighted regression to explore local crime patterns. Social Science Computer Review, 25(2), 174 – 193

Cai, Q. (2017). Integrating the macroscopic and microscopic traffic safety analysis using hierarchical models.

Cambridge Systematics, C., and Consult, A. E. C. O. M. (2007). A Recommended Approach to Delineating TAZ in Florida. Prepared for: Federal Highway Administration Report, Washington, DC.

Carrick, G., Bejleri, I., and Ouyang, Y. (2014). Methodological approach to spatiotemporal optimization of rural freeway enforcement in Florida. Transportation Research Record, 2425(1), 1-9.

Carvalho, J. R., and Lavor, S. C. (2008). Repeat property criminal victimization and income inequality in Brazil. Revista Economi A, 9, 87-110.

Charlton M., Fotheringham A.S., (2009). Geographically weighted regression. White paper. National Centre for Geocomputation, National University of Ireland Maynooth, Maynooth, Co Kildare, Ireland

Chasco, C., García, I., and Vicéns, J. (2007). Modeling spatial variations in household disposable income with geographically weighted regression.

Chatterjee, S., and Hadi, A. S. (2015). Regression analysis by example. John Wiley and Sons.

Chen, F., Ma, X., and Chen, S. (2016). Investigation of interaction between traffic safety, law enforcement and environment (No. MPC 16-311). Mountain Plains Consortium.

Cheng, W., and Washington, S. (2008). New criteria for evaluating methods of identifying hot spots. Transportation Research Record, 2083(1), 76-85.

Cheng, W., and Washington, S. P. (2005). Experimental evaluation of hotspot identification methods. Accident Analysis and Prevention, 37(5), 870-881.

Cheung, C., Shalaby, A., Persaud, B., and Hadayeghi, A. (2008). Models for safety analysis of road surface transit. Transportation Research Record: Journal of the Transportation Research Board, (2063), 168-175.

Chi, G., and Zhu, J. (2008). Spatial regression models for demographic analysis. Population Research and Policy Review, 27(1), 17-42.

Cho, S. H., Lambert, D. M., and Chen, Z. (2010). Geographically weighted regression bandwidth selection and spatial autocorrelation: an empirical example using Chinese agriculture data. Applied Economics Letters, 17(8), 767-772.

Chung, K., and Ragland, D. R. (2007). Method for generating continuous risk profile for highway collisions (No. 07-2935).

Chung, K., and Ragland, D. R. (2009). The continuous risk profile approach for the identification of high collision concentration locations on congested highways. In Transportation and traffic theory 2009: golden jubilee (pp. 463-480). Springer, Boston, MA.

Chung, K., Jang, K., Madanat, S., and Washington, S. (2011). Proactive detection of high collision concentration locations on highways. Procedia-social and behavioural sciences, 17, 634-645.

Chung, K., Ragland, D.R., Madanat, S and Oh, S. (2009), The continuous risk profile approach for the identification of high collision concentration locations on congested highways. Proceeding of 19th ISTTT, 463-480.

Collins, K., Babyak, C., and Moloney, J. (2006). Treatment of spatial autocorrelation in geocoded crime data. Proceedings of the American Statistical Association Section on Survey Research Methods, 2864-2871.

Comber, A. J., Brunsdon, C., and Radburn, R. (2011). A spatial analysis of variations in health access: linking geography, socio-economic status and access perceptions. International journal of health geographic, 10(1), 44.

Conor, P. (2018). Police resources in Canada, 2017. Juristat: Canadian Centre for Justice Statistics, 1-25.

Cook, C., (2012). Implementation of an area traffic officer program. A Leadership White Paper Submitted in Partial Fulfillment Required for Graduation from the Leadership Command College the Bill Blackwood Law Enforcement Management Institute of Texas.

Curman, A. S., Andresen, M. A., and Brantingham, P. J. (2015). Crime and place: a longitudinal examination of street segment patterns in Vancouver, BC. Journal of Quantitative Criminology, 31(1), 127-147.

Da Silva, A. R., and Rodrigues, T. C. V. (2014). Geographically weighted negative binomial regression—incorporating overdispersion. Statistics and Computing, 24(5), 769-783.

Datta, T., Feber, D., Schattler, K., and Datta, S. (2000). Effective safety improvements through low-cost treatments. Transportation Research Record: Journal of the Transportation Research Board, (1734), 1-6.

Davis, C. E. (2007). Regression to the mean. Wiley Encyclopedia of Clinical Trials, 1-2.

De Pauw, E., Daniels, S., Brijs, T., Wets, G., and Hermans, E. (2013). The magnitude of the regression to the mean effect in traffic crashes. Transportation Research Board.

Deane, G., Messner, S. F., Stucky, T. D., McGeever, K., and Kubrin, C. E. (2008). Not 'islands, entire of themselves': Exploring the spatial context of city-level robbery rates. Journal of Quantitative Criminology, 24(4), 363-380.

Diefenbach, D. L., and West, M. D. (2001). Violent crime and Poisson regression: A measure and a method for cultivation analysis. Journal of Broadcasting and Electronic Media, 45(3), 432-445.

DiRienzo, C., Fackler, P., and Goodwin, B. K. (2000, August). Modeling spatial dependence and spatial heterogeneity in county yield forecasting models. In Proceedings of the American Agricultural Economics Association Annual Meeting, Tampa, FL, USA (Vol. 1).

Dong, C., Clarke, D. B., Yan, X., Khattak, A., and Huang, B., (2014). Multivariate random-parameters zero-inflated negative binomial regression model: An application to estimate crash frequencies at intersections. Accident Analysis and Prevention, 70, 320-329.

Drawve, G., Grubb, J., Steinman, H., and Belongie, M. (2018). Enhancing data-driven law enforcement efforts: exploring how risk terrain modeling and conjunctive analysis fit in a crime and traffic safety framework. American Journal of Criminal Justice, 1-19.

Du, H., and Mulley, C. (2012). Understanding spatial variations in the impact of accessibility on land value using geographically weighted regression. Journal of Transport and Land Use, 5(2), 46-59.

El-Basyouny, K., and Sayed, T. (2009). Collision prediction models using multivariate Poisson-lognormal regression. Accident Analysis and Prevention, 41(4), 820-828.

El-Basyouny, K., Barua, S., and Islam, M. T. (2014). Investigation of time and weather effects on crash types using full Bayesian multivariate Poisson lognormal models. Accident Analysis and Prevention, 73, 91-99.

Ellen, I. G., Lacoe, J., and Sharygin, C. A. (2013). Do foreclosures cause crime? Journal of Urban Economics, 74, 59-70.

Elvik, R. (2008). A survey of operational definitions of hazardous road locations in some European countries. Accident Analysis and Prevention, 40(6), 1830-1835.

Elvik, R., Vaa, T., Hoye, A., and Sorensen, M. (Eds.). (2009). The handbook of road safety measures. Emerald Group Publishing.

Fahrmeir, L., and Tutz, G. (2013). Multivariate statistical modelling based on generalized linear models. Springer Science and Business Media.

Fahrmeir, L., Kneib, T., Lang, S., and Marx, B. (2013). Regression: models, methods and applications. Springer Science and Business Media.

Farber, S., and Páez, A. (2007). A systematic investigation of cross-validation in GWR model estimation: empirical analysis and Monte Carlo simulations. Journal of Geographical Systems, 9(4), 371-396.

Farrington, D. P., and Welsh, B. C. (2006). How important is “regression to the mean” in area-based crime prevention research? Crime Prevention and Community Safety, 8(1), 50-60.

FDOT (2015). Highway safety manual user guide. Florida Department of Transportation, United States of America.

Fotheringham, A. S. (2009). The problem of spatial autocorrelation and local spatial statistics. Geographical analysis, 41(4), 398-403.

Fotheringham, A.S., C. Brunsdon, and Charlton M.E. (2002). Geographically weighted regression: The analysis of spatially varying relationships. Chichester: Wiley.

Freilich, J. D., Adamczyk, A., Chermak, S. M., Boyd, K. A., and Parkin, W. S. (2015). Investigating the applicability of macro-level criminology theory to terrorism: A county-level analysis. Journal of Quantitative Criminology, 31(3), 383-411.

Getis, A. (1999). Spatial statistics. Geographical information systems, 1, 239-251.

Getis, A. (2008). A history of the concept of spatial autocorrelation: A geographer's perspective. Geographical Analysis, 40(3), 297-309.

Getis, A., and Ord, J. K. (1992). The analysis of spatial association by use of distance statistics. Geographical analysis, 24(3), 189-206.

Geurts, K., and Wets, G. (2003). Black spot analysis methods: Literature review. Flemish Research Center for Traffic Safety, Diepenbeek, Belgium.

Giuffrè, O., Granà, A., Giuffrè, T., Marino, R., and Marino, S. (2014). Estimating the safety performance function for urban un-signalized four-legged one-way intersections in Palermo, Italy. Archives of Civil Engineering, vol. 60, No.1, pp. 41-54.

Gkritza, K., Souleyrette, R. R., Baird, M. J., and Danielson, B. J. (2013). Empirical Bayes approach for estimating urban deer-vehicle crashes using police and maintenance records. Journal of Transportation Engineering, 40(2), 04013002.

Gover, A. R., Pérez, D. M., and Jennings, W. G. (2008). Gender differences in factors contributing to institutional misconduct. The Prison Journal, 88(3), 378-403.

Grembek, O., Kim, K., Kwon, O. H., Lee, J., Liu, H., Park, M. J., and Madanat, S. M. (2012). Experimental evaluation of the continuous risk profile (CRP) approach to the current Caltrans methodology for high collision concentration location identification.

Griffith, D. A. (2008). Spatial-filtering-based contributions to a critique of geographically weighted regression (GWR). Environment and Planning A, 40(11), 2751-2769.

Haberman, C. P. (2017). Overlapping hot spots? Examination of the spatial heterogeneity of hot spots of different crime types. Criminology & Public Policy, 16(2), 633-660.

Haberman, C. P., and Ratcliffe, J. H. (2015). Testing for temporally differentiated relationships among potentially criminogenic places and census block street robbery counts. Criminology, 53(3), 457-483.

Hadayeghi, A., Shalaby, A. S., and Persaud, B. N. (2010). Development of planning level transportation safety tools using Geographically Weighted Poisson Regression. Accident Analysis and Prevention, 42(2), 676-688.

Hadayeghi, A., Shalaby, A. S., Persaud, B. N., and Cheung, C. (2006). Temporal transferability and updating of zonal level accident prediction models. Accident Analysis and Prevention, 38(3), 579-589.

Haining, R. P. (2009). Spatial autocorrelation and the quantitative revolution. Geographical Analysis, 41(4), 364-374.

Haining, R., Law, J., & Griffith, D. (2009). Modelling small area counts in the presence of overdispersion and spatial autocorrelation. Computational Statistics & Data Analysis, 53(8), 2923-2937.

Hauer, E. (1992). Empirical Bayes approach to the estimation of “unsafety”: the multivariate regression method. Accident Analysis and Prevention, 24(5), 457-477.

Hauer, E., Council, F., and Mohammedshah, Y. (2004). Safety models for urban four-lane undivided road segments. Transportation Research Record: Journal of the Transportation Research Board, (1897), 96-105.

Hauer, E., Harwood, D., Council, F., and Griffith, M. (2002). Estimating safety by the empirical Bayes method: a tutorial. Transportation Research Record: Journal of the Transportation Research Board, (1784), 126-131.

Hautzinger, H., Pastor, C., Pfeiffer, M., and Schmidt, J. (2007). Analysis methods for accident and injury risk studies. TRACE deliverable, 7.

Highway Safety Manual (2010). American Association of State Highway and Transportation Officials (AASHTO). ISBN: 9781560514770

Huang, B., Wu, B., and Barry, M. (2010). Geographically and temporally weighted regression for modeling spatio-temporal variation in house prices. International Journal of Geographical Information Science, 24(3), 383-401.

Huang, H., Abdel-Aty, M., and Darwiche, A. (2010). County-level crash risk analysis in Florida: Bayesian spatial modeling. Transportation Research Record: Journal of the Transportation Research Board, (2148), 27-37.

Huang, H., Song, B., Xu, P., Zeng, Q., Lee, J., and Abdel-Aty, M. (2016). Macro and micro models for zonal crash prediction with application in hot zones identification. Journal of transport geography, 54, 248-256.

Huang, H., Zhou, H., Wang, J., Chang, F., and Ma, M. (2017). A multivariate spatial model of crash frequency by transportation modes for urban intersections. Analytic methods in accident research, 14, 10-21.

Institute of Transportation Engineers (2009). Before-and-after study technical brief. Transportation Engineers.

Jacquez, G. M. (2010). Space-time intelligence system software for the analysis of complex systems. In Handbook of Applied Spatial Analysis (pp. 113-124). Springer, Berlin, Heidelberg.

Jeon, J. H., Kho, S. Y., Park, J. J., and Kim, D. K. (2012). Effects of spatial aggregation level on an urban transportation planning model. KSCE Journal of Civil Engineering, 16(5), 835-844.

Jiang, B. (2015). Geospatial analysis requires a different way of thinking: The problem of spatial heterogeneity. GeoJournal, 80(1), 1-13.

Kim, Y. A. (2018). Examining the relationship between the structural characteristics of place and crime by imputing census block data in street segments: Is the pain worth the gain? Journal of Quantitative Criminology, 34(1), 67-110.

Klinger, D., Rosenfeld, R., Isom, D., and Deckard, M. (2016). Race, crime, and the micro-ecology of deadly force. Criminology and Public Policy, 15(1), 193-222.

Kondo, M., Hohl, B., Han, S., and Branas, C. (2016). Effects of greening and community reuse of vacant lots on crime. Urban studies, 53(15), 3279-3295.

Kononov, J., Bailey, B., and Allery, B. (2008). Relationships between safety and both congestion and number of lanes on urban freeways. Transportation Research Record: Journal of the Transportation Research Board, (2083), 26-39.

Kyratso, M., and Yiorgos, P. (2004, September). Defining a geographically weighted regression model of urban evolution. Application to the city of Volos, Greece. In 44th European Congress of the European Regional Science Association, University Of Porto, Porto, Portugal (August 25–29, 2004). Http://Www-Sre. Wu-Wien. Ac. At/Ersa/Ersaconfs/Ersa04/PDF/507. Pdf. Accessed (Vol. 7).

Ladron de Guevara, F., Washington, S., and Oh, J. (2004). Forecasting crashes at the planning level: simultaneous negative binomial crash model applied in Tucson, Arizona. Transportation Research Record: Journal of the Transportation Research Board, (1897), 191-199.

Lee, A. H., Stevenson, M. R., Wang, K., and Yau, K. K. (2002). Modeling young driver motor vehicle crashes: data with extra zeros. Accident Analysis and Prevention, 34(4), 515-521.

Lee, J., Abdel-Aty, M., and Jiang, X. (2015). Multivariate crash modeling for motor vehicle and non-motorized modes at the macroscopic level. Accident Analysis and Prevention, 78, 146-154.

Lee, J., and Abdel-Aty, M. (2018). Macro-level analysis of bicycle safety: Focusing on the characteristics of both crash location and residence. International Journal of Sustainable Transportation, 12(8), 553-560.

Lee, J., Chung, K., and Kang, S. (2016). Evaluating and addressing the effects of regression to the mean phenomenon in estimating collision frequencies on urban high collision concentration locations. Accident Analysis and Prevention, 97, 49-56.

Lee, J., Yasmin, S., Eluru, N., Abdel-Aty, M., and Cai, Q. (2018). Analysis of crash proportion by vehicle type at traffic analysis zone level: a mixed fractional split multinomial logit modeling approach with spatial effects. Accident Analysis and Prevention, 111, 12-22.

Lee, M. R., and Thomas, S. A. (2010). Civic community, population change, and violent crime in rural communities. Journal of Research in Crime and Delinquency, 47(1), 118-147.

Lee, M. T., Martinez, R., and Rosenfeld, R. (2001). Does immigration increase homicide? Negative evidence from three border cities. The Sociological Quarterly, 42(4), 559-580.

Lee, S. I. (2017). Correlation and spatial autocorrelation. In Encyclopedia of GIS (pp. 360-368). Springer International Publishing.

Lee, S., Kim, C., Kim, D. K., and Lee, C. (2013). Identifying hotspots on freeways using the continuous risk profile with hierarchical clustering analysis. Journal of Korean Society of Transportation, 31(4), 85-94.

Li Z., Wand W., Liu P., Bigham J.M., Ragland D.R., (2013)., Using geographical weighted Poisson regression for county level crash modeling. Safety Science 58, 89-97.

Li, R., El-Basyouny, K., and Kim, A. (2015). Before-and-after empirical Bayes evaluation of automated mobile speed enforcement on urban arterial roads. Transportation Research Record: Journal of the Transportation Research Board, (2516), 44-52.

Li, Y., and Zhu, K. (2015). Spatial dependence and heterogeneity in the location processes of new high-tech firms in Nanjing, China. Papers in Regional Science.

Li, Z., Knight, S., Cook, L. J., Hyde, L. K., Holubkov, R., and Olson, L. M. (2008). Modeling motor vehicle crashes for street racers using zero-inflated models. Accident Analysis and Prevention, 40(2), 835-839.

Li, Z., Lee, Y., Lee, S. H., Valiou, E., (2011). Geographically-weighted regression models for improved predictability of urban intersection vehicle crashes. In Transportation and Development Institute Congress 2011: Integrated Transportation and Development for a Better Tomorrow (pp. 1315-1329).

Light, M. T., and Harris, C. T. (2012). Race, space, and violence: exploring spatial dependence in structural covariates of white and black violent crime in US counties. Journal of Quantitative Criminology, 28(4), 559-586.

Lin, C. H., and Wen, T. H. (2011). Using geographically weighted regression (GWR) to explore spatial varying relationships of immature mosquitoes and human densities with the incidence of dengue. International journal of environmental research and public health, 8(7), 2798-2815.

Lloyd, C., Shuttleworth, I., (2005). Analysing commuting using local regression techniques: scale, sensitivity, and geographical patterning. Environment and Planning A, 37(1), 81-103.

Lo, C. C., and Zhong, H. (2006). Linking crime rates to relationship factors: The use of gender-specific data. Journal of Criminal Justice, 34(3), 317-329.

Loo, B. P., and Anderson, T. K. (2015). Spatial analysis methods of road traffic collisions. CRC Press.

Lopez, D., Glickman, M. E., Soumerai, S. B., and Hemenway, D. (2018). Identifying factors related to a hit-and-run after a vehicle-bicycle collision. Journal of Transport and Health, 8, 299-306.

Lovegrove, G. R., and Sayed, T. (2006). Using macrolevel collision prediction models in road safety planning applications. Transportation research record, 1950(1), 73-82.

Lovegrove, G. R., Sayed, T., (2006). Macro-level collision prediction models for evaluating neighbourhood traffic safety. Canadian Journal of Civil Engineering, 33(5), 609-621.

Lovegrove, G., and Sayed, T. (2007). Macrolevel collision prediction models to enhance traditional reactive road safety improvement programs. Transportation Research Record: Journal of the Transportation Research Board, (2019), 65-73.

Lovegrove, G., Lim, C., and Sayed, T. (2010). Community-based, macrolevel collision prediction model use with a regional transportation plan. Journal of Transportation Engineering, 2, 120-128.

Maggard Jr, D. L., and Jung, D. (2009). Irvine's area traffic officer program. The Police Chief, 77(3), 46-50.

Maher, M. J., and Summersgill, I. (1996). A comprehensive methodology for the fitting of predictive accident models. Accident Analysis and Prevention, 28(3), 281-296.

Maher, M., and Mountain, L. (2009). The sensitivity of estimates of regression to the mean. Accident Analysis and Prevention, 41(4), 861-868.

Maistros, A., and Schneider IV, W. H. (2018). A comparison of overtime patrol stops made inside and out of cluster identified hotspots. Traffic injury prevention, 19(3), 235-240.

Manan, M. M. A., Jonsson, T., and Várhelyi, A. (2013). Development of a Safety Performance Function for Motorcycle Accident Fatalities on Malaysian Primary Roads. Safety Science, 60, 13-20.

Marchant, P. R. (2004). A demonstration that the claim that brighter lighting reduces crime is unfounded. British Journal of Criminology, 44(3), 441-447.

Marchant, P., and Hall, C. (2007). Are the claims of lighting benefits true? How can we tell? Draft: October 2, 2007, 39.

Martinez, L. M., Dupont-Kieffer, A., and Viegas, J. M. (2010, July). An integrated application of zoning for mobility analysis and planning: the case of Paris Region. In 12th World Conference on Transport Research (p. 25p).

McClure, D., Levy, J., La Vigne, N., and Hayeslip, D. (2014). DDACTS evaluability assessment: Final report on individual and cross-site findings. Washington, DC: Justice Policy Center at the Urban Institute.

McGarrell E.F., Rydberg J., and Norris A., (2014). Flint DDACTS pilot evaluation: summary of findings. Michigan Justice Statistics Center, School of Criminal Justice Michigan State University, United States

McMillen, D. P. (2012). Quantile regression for spatial data. Springer Science and Business Media.

Medury, A., and Grembek, O. (2016). Dynamic programming-based hot spot identification approach for pedestrian crashes. Accident Analysis and Prevention, 93, 198-206.

Meng, Q. (2014). Regression kriging versus geographically weighted regression for spatial interpolation. International Journal of Advanced Remote Sensing and GIS, 3(1), pp-606.

Mense, A. T. (2001). Introduction to Regression Techniques. A doctoral dissertation submitted to RMS University in fulfilment for the Award of Doctor of Philosophy Degree, RMS University of Chicago, USA.

Metaxatos, P., and Center, U. T. (2015). Some Thoughts on the Use of TAZ in Transportation Planning.

Miaou, S. P., Hu, P. S., Wright, T., Rathi, A. K., and Davis, S. C. (1992). Relationship between truck accidents and highway geometric design: a Poisson regression approach. Transportation Research Record, (1376).

Montella, A. (2010). A comparative analysis of hotspot identification methods. Accident Analysis and Prevention, 42(2), 571-581.

Montgomery, D. C., Peck, E. A., and Vining, G. G. (2012). Introduction to linear regression analysis (Vol. 821). John Wiley and Sons.

Naderan A, And Shahi J. (2010). Aggregated crash prediction models: introducing crash generation concept. Accident Analysis and Prevention, 42(1), 339-346

Nakaya, T., Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2005). Geographically weighted Poisson regression for disease association mapping. Statistics in Medicine, 24(17), 2695-2717

National Highway Traffic Safety Administration (NHTSA). (2009). Data-driven approaches to crime and traffic safety (DDACTS): operational guidelines. Washington, DC: US Department of Transportation.

Noresah, M. S., and Ruslan, R. (2009, July). Modelling urban spatial structure using geographically weighted regression. In 18th World IMACS congress and MODSIM09 international congress on modelling and simulation, The Australian National University Canberra, ACT.

Oh, S., Chung, K., Ragland, D. R., and Chan, C. Y. (2009). Analysis of wet weather-related collision concentration locations: empirical assessment of continuous risk profile.

Osama, A., and Sayed, T. (2016). Evaluating the impact of bike network indicators on cyclist safety using macro-level collision prediction models. Accident Analysis and Prevention, 97, 28-37.

Páez, A., and Scott, D. M. (2005). Spatial statistics for urban analysis: a review of techniques with examples. GeoJournal, 61(1), 53-67.

Páez, A., Farber, S., and Wheeler, D. (2011). A simulation-based study of geographically weighted regression as a method for investigating spatially varying relationships. Environment and Planning A, 43(12), 2992-3010.

Park, H. H., Oh, G. S., and Paek, S. Y. (2012). Measuring the crime displacement and diffusion of benefit effects of open-street CCTV in South Korea. International Journal of Law, Crime and Justice, 40(3), 179-191.

Persaud, B. N., Retting, R. A., and Lyon, C. A. (2004). Crash reduction following installation of centerline rumble strips on rural two-lane roads. Accident Analysis and Prevention, 36(6), 1073-1079.

Persaud, B., and Lyon, C. (2007). Empirical Bayes before–after safety studies: lessons learned from two decades of experience and future directions. Accident Analysis and Prevention, 39(3), 546-555.

Pirdavani, A., Bellemans, T., Brijs, T., and Wets, G. (2014a). Application of geographically weighted regression technique in spatial analysis of fatal and injury crashes. Journal of Transportation Engineering, 140(8), 04014032.

Pirdavani, A., Bellemans, T., Brijs, T., Kochan, B., and Wets, G. (2014b). Assessing the road safety impacts of a teleworking policy by means of geographically weighted regression method. Journal of transport geography, 39, 96-110.

Pirdavani, A., Brijs, T., Bellemans, T., Kochan, B., and Wets, G. (2012). Application of different exposure measures in development of planning-level zonal crash prediction models. Transportation Research Record: Journal of the Transportation Research Board, (2280), 145-153.

Piza, E. L., Caplan, J. M., and Kennedy, L. W. (2014). Analyzing the influence of micro-level factors on CCTV camera effect. Journal of Quantitative Criminology, 30(2), 237-264.

Plassmann, F., and Lott, J. R. (2002). More readers of gun magazines, but not more crimes. Social Science Research Network electronic library, 2002(online), ID-320107.

Plümper, T., and Neumayer, E. (2010). Model specification in the analysis of spatial dependence. European Journal of Political Research, 49(3), 418-442.

Pridemore, W. A., and Grubesic, T. H. (2012). Community organization moderates the effect of alcohol outlet density on violence. The British journal of sociology, 63(4), 680-703.

Pulugurtha, S. S., and Sambhara, V. R. (2011). Pedestrian crash estimation models for signalized intersections. Accident Analysis and Prevention, 43(1), 439-446.

Pulugurtha, S. S., Duddu, V. R., and Kotagiri, Y. (2013). Traffic analysis zone level crash estimation models based on land use characteristics. Accident Analysis and Prevention, 50, 678-687.

Quddus, M. A. (2008). Modelling area-wide count outcomes with spatial correlation and heterogeneity: an analysis of London crash data. Accident Analysis and Prevention, 40(4), 1486-1497.

Radil (2016). Spatial analysis of crimes. The Handbook of Measurement Issues in Criminology and Criminal Justice, First Edition. Edited by Beth M. Huebner and Timothy S. Bynum. John Wiley and Sons

Ratcliffe, J. (2010). Crime mapping: spatial and temporal challenges. In Handbook of quantitative criminology (pp. 5-24). Springer, New York, NY.

Rhee, K. A., Kim, J. K., Lee, Y. I., and Ulfarsson, G. F. (2016). Spatial regression analysis of traffic crashes in Seoul. Accident Analysis and Prevention, 91, 190-199.

Rodegerdts, L. A., Nevers, B., Robinson, B., Ringert, J., Koonce, P., Bansen, J., and Neuman, T. (2004). Signalized intersections: informational guide (No. FHWA-HRT-04-091).

Rolison, Jonathan J., and Salissou Moutari (2017). Risk-exposure density and mileage bias in crash risk for older drivers. American journal of epidemiology 187(1), 53-59.

Rosser, G., Davies, T., Bowers, K. J., Johnson, S. D., and Cheng, T. (2017). Predictive crime mapping: arbitrary grids or street networks? Journal of quantitative criminology, 33(3), 569-594.

Sampson, R. J., and Raudenbush, S. W. (2001). Disorder in urban neighborhoods: Does it lead to crime. Washington, DC: US Department of Justice, Office of Justice Programs, National Institute of Justice.

Senn, S. (2011). Francis Galton and regression to the mean. Significance, 8(3), 124-126.

Shaheed, M. S., Gkritza, K., Hallmark, S. L., and Knapp, K. K. (2015). An application of the empirical Bayes method for identifying winter weather crash hot spots. In Transportation Research Board 94th Annual Meeting, No. 15-4785.

Shi, H., Laurent, E. J., LeBouton, J., Racevskis, L., Hall, K. R., Donovan, M., ... and Liu, J. (2006). Local spatial modeling of white-tailed deer distribution. Ecological Modelling, 190(1-2), 171-189.

Simes, J. T. (2017). Place and punishment: The spatial context of mass incarceration. Journal of Quantitative Criminology, 1-21.

Soltani, A., and Askari, S. (2017). Exploring spatial autocorrelation of traffic crashes based on severity. Injury, 48(3), 637-647.

Sporer, K., Anderson, A. L., and Peterson, J. (2017). Macro-and Micro-approaches to Crime Prevention and Intervention Programs. In Preventing Crime and Violence (pp. 169-176). Springer, Cham.

Srinivasan, B., Lyon, C., Persaud, B. N., Martell, C., and Baek, J. (2011). Methods for identifying high collision concentration locations (HCCL) for potential safety improvements: Phase II, Evaluation of alternative methods for identifying HCCL. Caltrans, Sacramento, CA: California Department of Transportation.

Statistics Canada (2011). Overview of the Census. Catalogue no. 98 301 XIE. http://www12.statcan.gc.ca/census-recensement/2011/ref/overview-apercu/pop9-eng.cfm

Steenbeek, W., and Weisburd, D. (2016). Where the action is in crime? An examination of variability of crime across different spatial units in The Hague, 2001–2009. Journal of Quantitative Criminology, 32(3), 449-469.

Stucky, T. D., Ottensmann, J. R., and Payton, S. B. (2012). The effect of foreclosures on crime in Indianapolis, 2003–2008. Social Science Quarterly, 93(3), 602-624.

Su, S., Lei, C., Li, A., Pi, J., and Cai, Z. (2017). Coverage inequality and quality of volunteered geographic features in Chinese cities: Analyzing the associated local characteristics using geographically weighted regression. Applied geography, 78, 78-93.

Takyi, E. A., Oluwajana, S. D., and Park, P. Y. (2018). Development of macro-level crime and collision prediction models to support data-driven approach to crime and traffic safety (DDACTS). Transportation Research Record, 0361198118777356.

Thapa R.B and Estoque R.C (2012). Geographically weighted Regression. In Geospatial Analysis in Progress in geospatial analysis. Murayama, Y. (Ed). Springer Science and Business Media.

The city of Regina (2016). Managing progress: doing what matters most. ftp://ftp.regina.ca/web_files/budget/2016budget.pdf

Timofeeva, A., and Tesselkina, K. (2016, October). Local spatial interaction modelling of graduate flows. In AIP Conference Proceedings., 1772(1), AIP Publishing.

Townsley, M., Birks, D., Ruiter, S., Bernasco, W., and White, G. (2016). Target selection models with preference variation between offenders. Journal of quantitative criminology, 32(2), 283-304.

Uyanık, G. K., and Güler, N. (2013). A study on multiple linear regression analysis. Procedia-Social and Behavioural Sciences, 106, 234-240.

Vandeviver, C., and Steenbeek, W. (2017). The (in) stability of residential burglary patterns on street segments: the case of Antwerp, Belgium 2005–2016. Journal of Quantitative Criminology, 1-23.

Wallace D., and Scott C., (2017). Neighborhood disorder. DOI: 10.1093/OBO/9780195396607-0154

Walters, G. D. (2007). Using Poisson class regression to analyze count data in correctional and forensic psychology: A relatively old solution to a relatively new problem. Criminal Justice and Behaviour, 34(12), 1659-1674.

Wang, J. H., Abdel-Aty, M. A., Park, J., Lee, C., and Kuo, P. F. (2015). Estimating safety performance trends over time for treatments at intersections in Florida. Accident Analysis and Prevention, 80, 37-47.

Wang, K., Ivan, J. N., Ravishanker, N., and Jackson, E. (2017). Multivariate Poisson lognormal modeling of crashes by type and severity on rural two-lane highways. Accident Analysis and Prevention, 99, 6-19.

Wang, N., Mei, C. L., and Yan, X. D. (2008). Local linear estimation of spatially varying coefficient models: an improvement on the geographically weighted regression technique. Environment and Planning A, 40(4), 986-1005.

Wang, X., Yang, J., Lee, C., Ji, Z., and You, S. (2016). Macro-level safety analysis of pedestrian crashes in Shanghai, China. Accident Analysis and Prevention, 96, 12-21.

Wang, Y., and Kockelman, K. M. (2013). A Poisson-lognormal conditional-autoregressive model for multivariate spatial analysis of pedestrian crash counts across neighbourhoods. Accident Analysis and Prevention, 60, 71-84.

Ward, M. D., and Gleditsch, K. S. (2018). Spatial regression models (Vol. 155). Sage Publications.

Washington, S., Karlaftis, M. G., and Mannering, F. L. (2003). Statistical and econometric methods for transportation data analysis (pp. 241-255). Boca Raton, FL: Chapman and Hall/CRC.

Wei, F., Alam, A., and Lovegrove, G. (2011). Macro-Level Collision Prediction Models Related to Bicycle Use. In ICTIS 2011: Multimodal Approach to Sustained Transportation System Development: Information, Technology, Implementation (pp. 1312-1323).

Wei, F., and Lovegrove, G. (2013). An empirical tool to evaluate the safety of cyclists: Community based, macro-level collision prediction models using negative binomial regression. Accident Analysis and Prevention, 61, 129-137.

Weisburd, D. (2008). Place-Based Policing (Ideas in American Policing Series, Number 9). Washington, DC: Police Foundation.

Wells, L. E., and Weisheit, R. A. (2012). Explaining crime in metropolitan and non-metropolitan communities. International Journal of Rural Criminology. 1, (2), 153 – 183.

Weslaco Police Department, Texas USA (2018). About Data-Driven Approaches to Crime and Traffic Safety. http://www.weslacotx.gov/departments/police?sub=crime-fusion-center-police

Wheeler D.C., Paez A., (2010). Geographically weighted regression. Handbook of applied spatial analysis: software tools, methods and applications. Fischer, M. M., and Getis, A. (Eds.). Springer Science and Business Media.

Whigham, P and Hay, G (2007). A Preliminary Investigation of the Stability of Geographically-Weighted Regression. The 19th Annual Colloquium of the Spatial Information Research Centre (SIRC) University of Otago, Dunedin, New Zealand December 6th-7th, 2007.

Windle, M. J., Rose, G. A., Devillers, R., and Fortin, M. J. (2009). Exploring spatial non-stationarity of fisheries survey data using geographically weighted regression (GWR): an example from the Northwest Atlantic. ICES Journal of Marine Science: Journal du Conseil, fsp224.

Winkelmann, R. (2008). Econometric analysis of count data. Springer Science and Business Media

Xu, P., Huang, H., and Dong, N. (2016). The modifiable areal unit problem in traffic safety: basic issue, potential solutions and future research. Journal of Traffic and Transportation Engineering (English Edition).

Xu, P., Huang, H., Dong, N., and Abdel-Aty, M. (2014). Sensitivity analysis in the context of regional safety modeling: identifying and assessing the modifiable areal unit problem. Accident Analysis and Prevention, 70, 110-120.

Yang, B. Z., and Loo, B. P. (2016). Land use and traffic collisions: A link-attribute analysis using Empirical Bayes method. Accident Analysis and Prevention, 95, 236-249.

Yang, S. M. (2010). Assessing the spatial–temporal relationship between disorder and violence. Journal of Quantitative Criminology, 26(1), 139-163.

Zegras, C. (2010). The built environment and motor vehicle ownership and use: Evidence from Santiago de Chile. Urban Studies, 47(8), 1793-1817.

Zhai, X., Huang, H., Gao, M., Dong, N., and Sze, N. N. (2018). Boundary crash data assignment in zonal safety analysis: an iterative approach based on data augmentation and Bayesian spatial model. Accident Analysis and Prevention, 121, 231-237.

Zhang, C., and Ivan, J. (2005). Effects of geometric characteristics on head-on crash incidence on two-lane roads in Connecticut. Transportation Research Record: Journal of the Transportation Research Board, (1908), 159-164.

Zhang, L., Ma, Z., and Guo, L. (2009). An evaluation of spatial autocorrelation and heterogeneity in the residuals of six regression models. Forest Science, 55(6), 533-548.

Zhang, Y., Bigham, J., Ragland, D., and Chen, X. (2015). Investigating the associations between road network structure and non-motorist accidents. Journal of transport geography, 42, 34-47.

Zhao, S., Khattak, A. J., and Thompson, E. C. (2015, March). Safety and Economic Assessment of Converting Two-Way Stop-Controlled Intersections to Roundabouts on High Speed Rural Highways. In Journal of the Transportation Research Forum (Vol. 54, No. 1).

Zhen, Z., Li, F., Liu, Z., Liu, C., Zhao, Y., Ma, Z., and Zhang, L. (2013). Geographically local modeling of occurrence, count, and volume of Downwood in Northeast China. Applied Geography, 37, 114-126

Zheng, L., Robinson, R. M., Khattak, A., and Wang, X. (2011). All accidents are not equal: using geographically weighted regressions models to assess and forecast accident impacts. In 3rd International Conference on Road Safety and Simulation.

CHAPTER 3:RESEARCH DATA AND MODELLING METHODS

This chapter provides information on the study area and sources of data used in this research. The chapter also discusses the method used to aggregate the data into areal units (TAZs). The treatment of boundary incidences (crimes and collision falling on the boundary of two or more spatial units) is discussed. The underlying assumptions used in the existing literature for boundary collision aggregation into spatial units are explained. The selected approach, i.e. the pre-weighting technique, for assigning boundary crimes and collisions into a spatial unit is highlighted. The research approach, modelling techniques, and goodness of fit tests used to identify best performing models are also discussed.

3.1 Study Area and Data

This research used City of Regina, Saskatchewan data to develop macro-level crime and collision prediction models. Regina is the capital city of the province of Saskatchewan and is in the prairie region of Canada. Over the years, Regina has grown in population (Statistics Canada, 2017a; Statistics Canada, 2017b; City of Regina, 2018) and social issues have also increased. The city has witnessed an unprecedented rise in the number of crimes and collisions. In the annual comparison of crime severity index among cosmopolitan areas of Canada, the city was ranked higher than any other of the other cities included. This year is one the reference years included in this study (Boyce et al., 2014). Figure 3-1 shows the crime severity index in Canada among comparison cosmopolitan cities in year 2013.

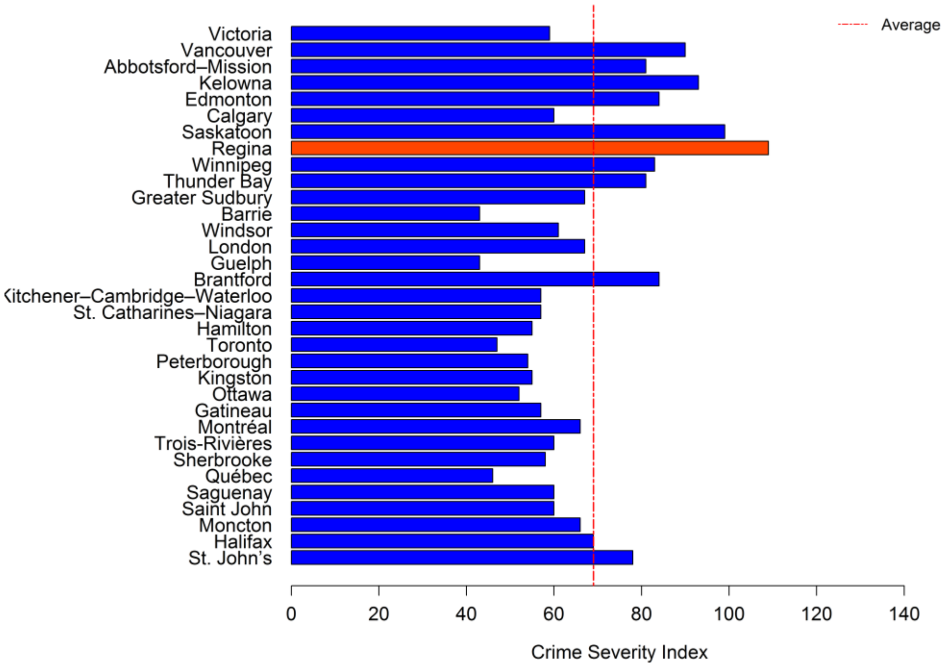


Figure 3-1: Crime Severity Index in Canada

(Statistics Canada, 2013)

In 2013, the city recorded a crime severity index of 109.3 which is measured as the total number of crimes per hundred thousand population. Regina's crime severity index was reported to be more than the national average of 68.7 and Regina still ranks high among other municipalities. In 2013, Saskatchewan Government Insurance, an agency responsible for the collection of traffic collision records for the province of Saskatchewan reported that the city ranked high in the number of fatalities recorded on roadways. Around this same period, the city ranked second to the city of Saskatoon in the number of total, injury and property damage only collisions. The simultaneous occurrence of crime and collision in Regina made it a suitable choice for this study.

3.2 Data Collection

The dataset used in this research was obtained from various government agencies in Regina. These agencies include the City of Regina (COR), Saskatchewan Government Insurance (SGI) and Regina Police Service (RPS). The data obtained from these agencies includes crime, collision, roadway inventory and demographic data. Land use data (land use classification information) was obtained from an online open source platform known as open.regina.ca.

Descriptions of the databases are provided in the next section.

3.2.1 Crime Database

The RPS collects records for crimes occurring within the City of Regina. These are stored in a database along with the nature of the crimes (assault, robbery, murder etc). The crime location (with location coordinates) and time and characteristics of the offenders (age, gender, ethnicity, citizenship, etc) are also stored in this database.

The crime data provided by RPS were obtained in three text files: the occurrence file, the address file and the person file. The occurrence file provides information on the type of crime that was reported and the time of occurrence (start time, end time and time reported to the police). The address file contains the address where the crimes occurred and the geographic coordinates. The person file contains the record of person(s) involved. Common to the three files provided by RPS is the occurrence file number. The occurrence file number facilitated merging of the three files into a comprehensive crime database.

This research focused on only 5 years (2009-2013) of crime records. However, there were some data quality issue with the crime records during the study period, in particular the time of occurrence of the crimes. Some of the crimes in the database lacked the time of occurrences and some lacked the start time, end time or reported time.

The time when the crime was reported served as the basis for the analysis conducted in this research. Where times were missing, they could not be included in the analysis.

The list of crime types recorded in the database of the RPS are numerous. Based on consultation and recommendation by the city of RPS, 10 crime types were selected. These crimes were Arson, Assault, Break and Enter, Mischief, Sexual Assault, Murder, Robbery, Theft, Theft of Automobile and Theft from Automobile. These 10 types of crime were then categorized as violent, non-violent and total crimes. In this research, crimes that can result to bodily harm or death are classed as violent, otherwise they are classified as non-violent. The summation of the two categories gives the total crimes.

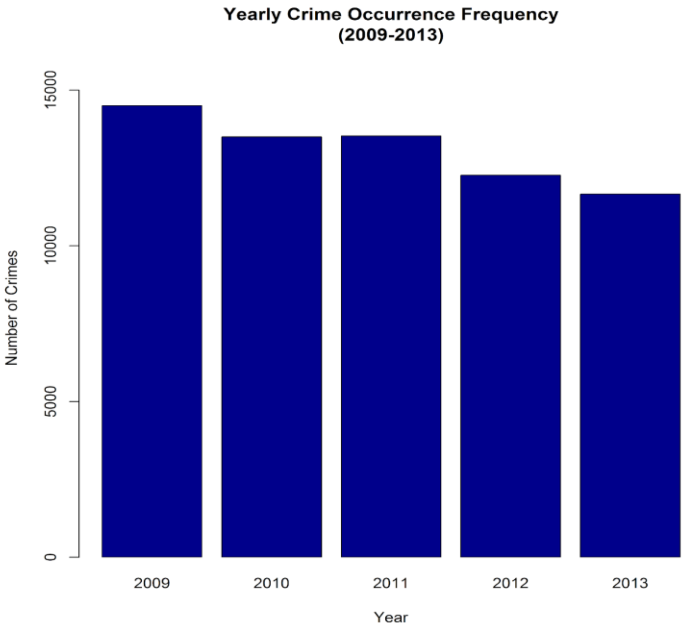


Figure 3-2: Yearly Distribution of Total of Ten Crimes in Regina

The total number of crimes has been on the downward trend over the past years.

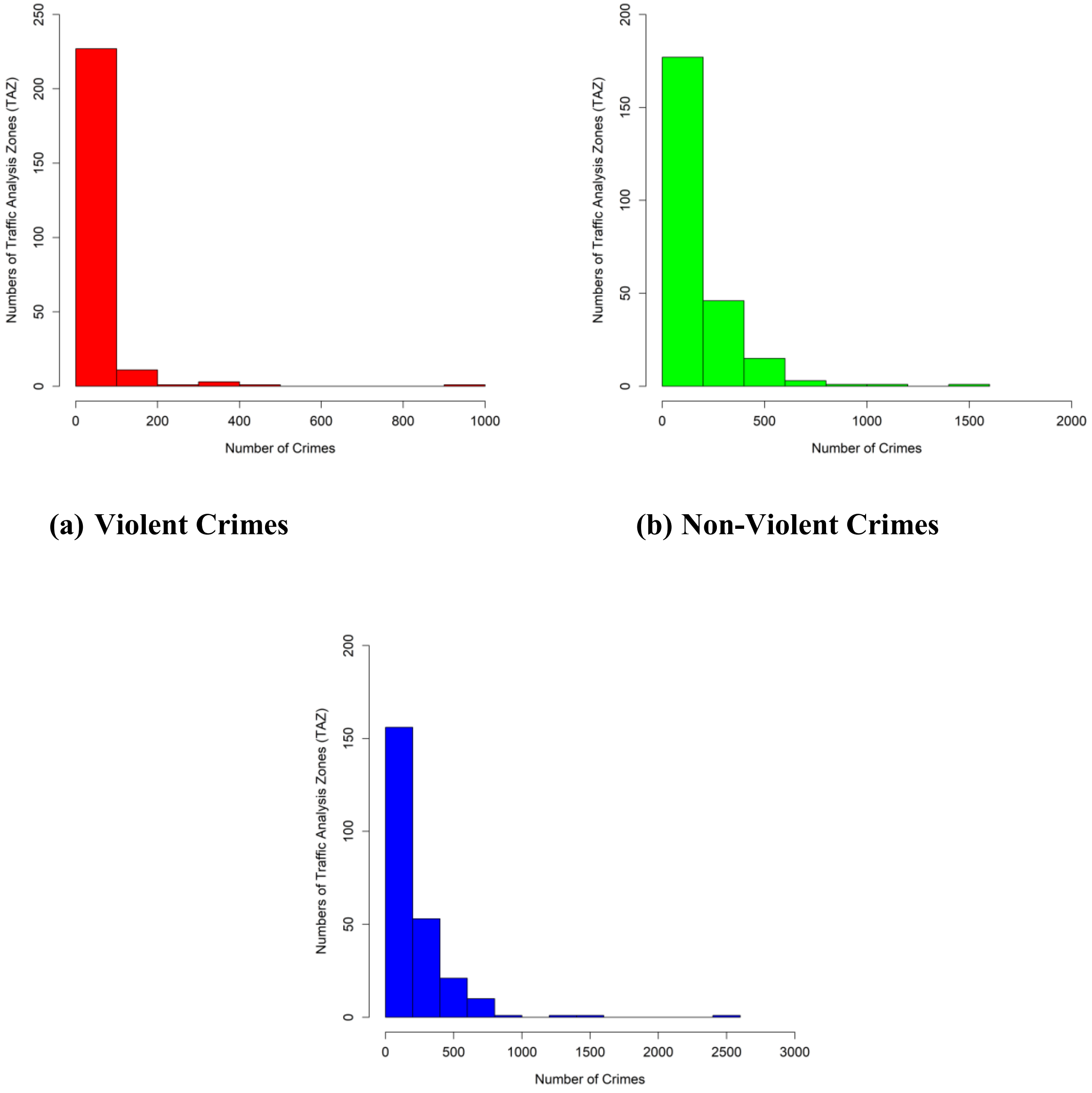
In 2009, about 15,000 crimes were recorded and this total decreased slightly in 2010 to about 14,000. The numbers of crimes in the year 2011 remained the same as 2010 but dropped in the years 2012 and 2013 (see Figure 3-2). During the 5 year study period, about 65,505 crimes were recorded in the Regina. Importantly, not all the coordinates of these crimes were available and some did not fall within the Regina city limit when the crimes were allocated to TAZs. Only crimes with known location coordinates were aggregated and assigned to the TAZs.

For macro-level analysis of crimes conducted in this research, a total of 50,284 crimes were used for further analysis. About 18% (9,181) of these crimes were violent and 82% (41,103) were non-violent.

About 1% percent (284) of the total violent crimes were Arson, 14% (6,922) were Assault, 3% (1,315) were Robbery, and 1% (636) were Sexual Assault. There were 24 murders, a very small percentage.

For non-violent crimes, Break and Enter constituted about 12% (6053), Mischief about 22% (10,908), Theft about 23% (11,436), Theft from Auto about 16% (8,057), and Theft of Auto about 9% (4,649).

The distribution of the three crime categories (violent, non-violent and total) aggregated into 244 Regina traffic analysis zone is shown on Figure 3-3.



(a) Violent Crimes

(b) Non-Violent Crimes

(c) Total Crimes

Figure 3-3: Distribution of Aggregated Regina Crimes

The distribution of Regina crimes across traffic zones shows skewness. They crimes are non-normally distributed.

3.2.2 Collision Database

SGI collects records of collisions occurring on roadways within the City of Regina and all other cities in the province of Saskatchewan, Canada. The information is stored in a database known as the Traffic Accident Information System (TAIS). According to SGI (2013), TAIS is a database on which details of traffic collisions occurring on roadways in the province of Saskatchewan are recorded. Collisions usually stored on these databases include those involving injury, fatalities and property damage only. The database also includes information on hit and run collisions and collisions involving impaired driving in which vehicles involved are towed and reported to the police.

The advantage of TAIS is that it provides detailed information on collisions to many agencies that might be interested in assessing transportation safety initiatives. The gathering of details related to traffic collisions recorded on this database is done in collaboration with law enforcement agencies and the staff of SGI who work to investigate and record collision occurrences on TAIS. Aside from the types of collision recorded in this database, information is provided on the characteristics of roadways on which they occur. This could be intersections or roadway segments. The street address details and their UGRID (the location identifier for segments and intersections on which the collisions occur) are provided in TAIS. Generally, in Saskatchewan, collisions that occur on roadway elements are described using the UGRID.

Collision records obtained for this research were provided at a disaggregated level and not in aggregated form.

The traffic collisions data obtained from the SGI were provided in three separate files (tables): collision, vehicle and occupant. Each of these databases contained a common identifier.

The collision record contain records relating to the nature of the collisions and the circumstances that lead to their occurrence. Information recorded include location details using the UGRID, date and time of occurrence, severity (fatal, injury or property damage only), collision cost and weather conditions at the time of the collision. The collisions are recorded by type of road element (intersection or roadway segment) on which they occurred. The vehicle table contains information about the number of vehicles and details of the driver (age, gender, date of birth, etc) involved in the collision. Also contained in the vehicle table are the contributing factors to the collision. The occupant table contains the record of persons in the vehicle at the time of collision. Using the UGRID, the tables were merged and two separate collision databases were created: intersection collisions and roadway segment collisions created ready for subsequent aggregation analysis.

The databases provided by SGI included several years of information relating to collisions that occurred within the City of Regina. However, only five years (2009-2013) of collisions were used for further analysis. The analysis considered three collision severity levels: fatal-injury, property damage only and total. Fatal-injury collisions are those that result to injuries to persons or loss of life. In principle, the number of fatal collisions is always few so they are usually combined with injury collisions for analysis purposes. Property damage only collisions refer to collisions in which property such as vehicles or highway infrastructure is damaged, but no one is injured or killed (Blincoe et al., 2002, Highway Safety Manual, 2010). The sum of the fatal-injury and property damage only collisions is the total collisions.

Figure 3-4 shows the number of collisions that occurred within the five year study period considered in this research. In 2009, the City of Regina experienced a total of 6,513 collisions. This number decreased in 2010 and 2011 compared with the year 2009. An increasing trend was seen in the year 2012 when referenced to the years 2010 and 2011 (collisions increased to 5,944 in 2012). This then dropped to 5,787 in 2013 (see Figure 3-4).

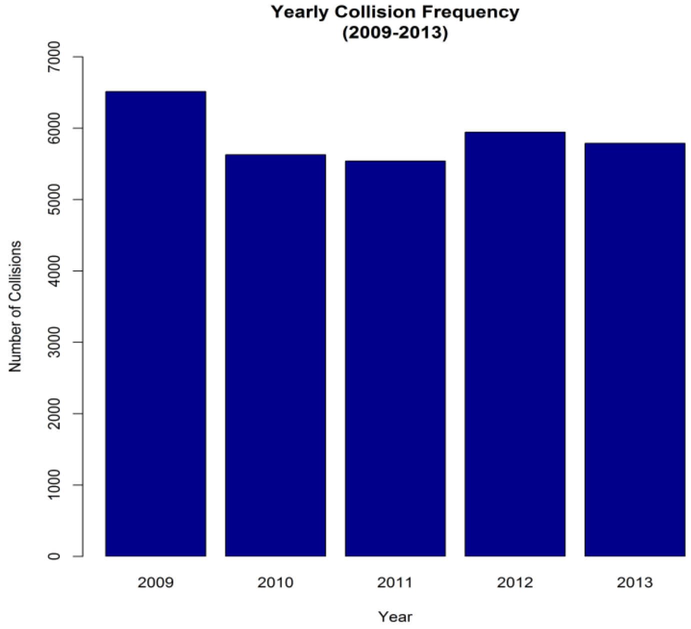
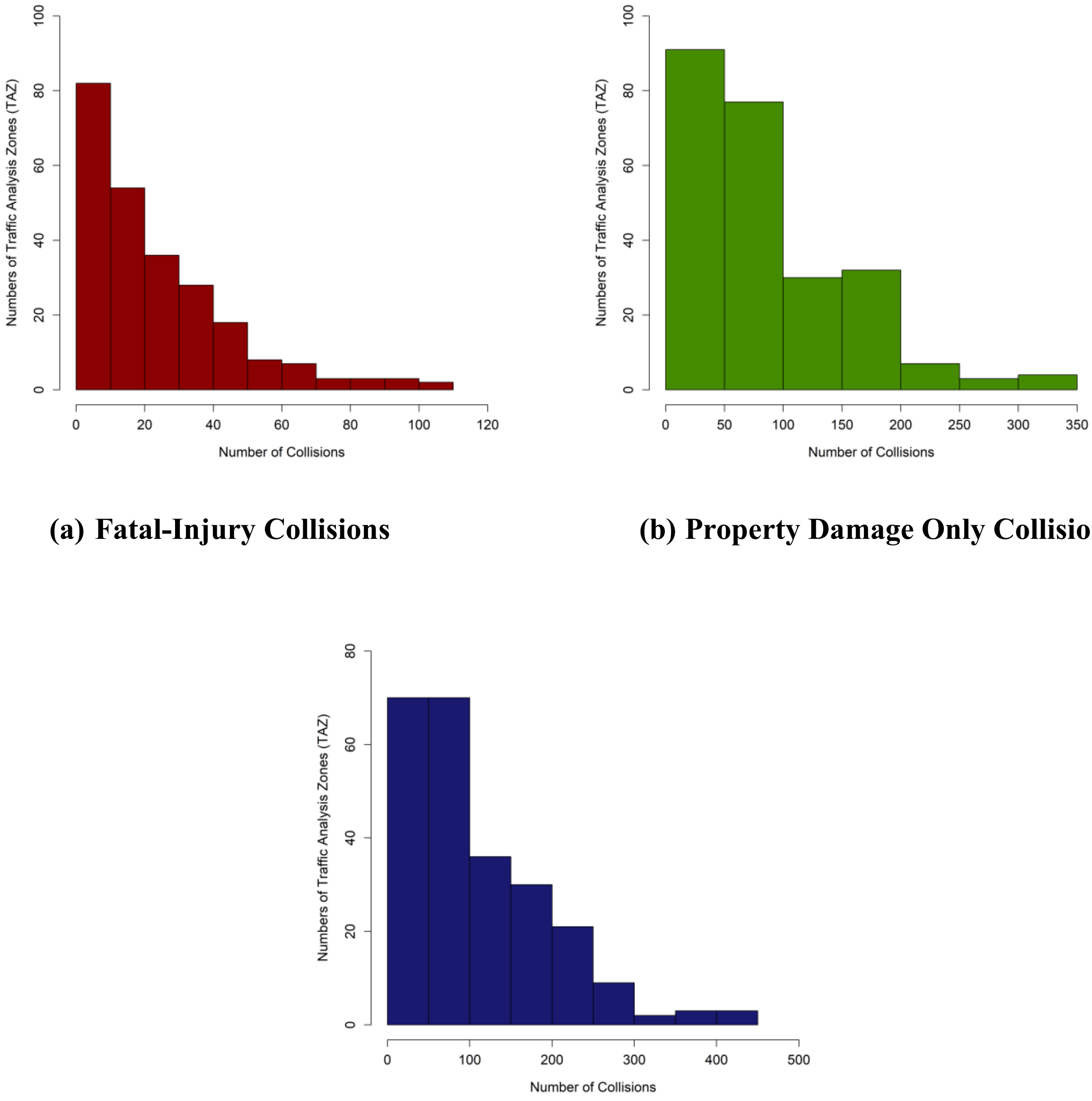


Figure 3-4: Yearly Distribution of Collisions in Regina

During the five year study period, the City of Regina recorded 29,411 collisions. About 20% (6,010) of these collisions were fatal-injury while 80% (23,366) were property damage only.

By using the UGRID and coordinates generated for each collision, the collisions were assigned to city TAZs. However, there were boundary issues, i.e., collisions falling at the boundary of two or more zones. Also, not all the collisions recorded has UGRID identifier. These collisions could not be considered.

Aggregation of the collisions into the 244 TAZs of Regina resulted in a total of 26,642 collisions with 21.7% (5,759) being fatal-injury collisions and 78.3% (20,883) being property damage only collisions. See Figure 3-5 for distribution of collisions across zones. The methods of aggregation and the process for dealing with collisions that fell on a boundary are discussed in subsequent sections.



(a) Fatal-Injury Collisions

(b) Property Damage Only Collisions

(c) Total Collisions

Figure 3-5: Distribution of Aggregated Regina Collisions

Similar to the crime database, the distribution of Regina collisions aggregated into zones is skewed.

3.2.3 Sociodemographic Database

The demographic database of Regina was provided by the City of Regina. This data was obtained in aggregated form, i.e., at the TAZs level. In most cases, sociodemographic data are provided at an aggregated level due to confidentiality constraints.

The demographic database obtained from City of Regina contains the records by age category, proportion of persons enrolled in school, and total number of persons per TAZ. The database also includes land use information (office, retail, etc.) in a ready to use format.

3.2.4 Roadway Inventory

Roadway inventory information was supplied by SGI in ArcGIS shape file format. It included details of all road networks within the City of Regina.

The shape file provided information about road types by functional class, use and speed. It also provided information about the location identifier known as the UGRID of these roads. Similarly, the intersection shape file for the City of Regina was provided. It also contains the UGRID location identifiers as one of the attributes. However, the Annual Average Daily Traffic (AADT) of all road segments in the City of Regina was not measured. Only some roads have their AADT provided. In cases where this are not measured or available, the average of AADT by road functional category (local, collectors, arterial etc.) was assigned to each road with missing information based on functional category. Using the AADT provided and the assigned average, Vehicle Kilometer Travelled (VKMT) was calculated as the product of their AADT and length). This approach to estimating VKMT was applied to all the roads in the City of Regina.

3.2.5 Land use

The land use information described in this research characterizes how the land area has been assigned for use.

In land use shape files of Regina obtained from open.regina.ca, land areas are classified as commercial, industrial, residential (low, medium and high) density, recreational or open space, railway, and airport area. The area per TAZ of each land use was determined in ArcGIS by overlaying the land shapefile on the City of Regina TAZ. The area of each land use was then estimated.

3.3 Data Aggregation

Some data obtained for this study were provided at disaggregated level, i.e. they were not provided as aggregated information in areal units.

As crime data obtained from the City of Regina police service were reported using coordinates of the location where they occur, aggregation of the crime data into the TAZs was quite straightforward. However, before importing the crime database and crime coordinates into the ArcGIS environment, the 10 types of crimes which this research focuses on were identified and approved by Regina police authorities. After this approval, the crime database was imported into ArcGIS and the coordinates plotted considering the specified ten types of crimes. This was overlaid on the TAZ shape file to determine the number of crimes within each zone (within the 2013 city limits). The sum of each crime type for each zone was then determined.

The collision data provided the UGRID, but not the coordinates. Intersection collisions and roadway segment collisions were provided in separate files. Road segment in Regina were converted into point shape files and the coordinates generated. This information was exported into Microsoft Access and merged with the collision database using the common identifier, i.e., the UGRIDs. From the merging of two files, coordinates were generated for collisions occurring at segment level.

A similar approach was used for database containing intersection collisions. Intersection shape files were imported into ArcGIS and the coordinates generated. The information was exported to Microsoft Access and merged with collision records using the common identifier known as the UGRID.

The collision databases for intersections and road segments were combined into one database. Using the coordinate of all the collisions, the collisions were plotted in ArcGIS and spatially overlaid on the TAZs to determine the number of collisions by severity (fatal, injury and property damage only) within each TAZ.

The approach employed in the aggregation of collisions that occurred at segment level was employed in determining the total VKMT per TAZ. Since this information had been estimated prior to conversion of each segment into a points using the segment coordinates, it was quite straightforward to estimate the VKMT per zone. The coordinates of the segments were plotted and spatially overlaid on traffic zones. The sum of the VKMT of all the road segment centroids within each zone was assigned to the TAZ.

The same approach was used to determine each TAZ's total length of roads by functional class, use and speed.

To determine the numbers of legs per intersection, the number of legs was first determined in the ArcGIS environment. An intersection shape file was overlaid with the road segments in original form before conversion to points. Through this overlay, the number of legs per intersection was determined. Any intersection point that recorded one or two legs was removed from the analysis. This is because these intersections represent dead ends streets that are mostly represented as point in ArcGIS digitization. Similarly, two leg intersections do not exist. Two leg intersections appearing in the database could be as a result of a digitization error causing a breakage in a continuous road segment.

After determining the number of legs per intersection, the information was overlaid on the TAZs. Summaries of intersections by numbers of legs and total for each TAZ were obtained. However, in this analysis, weighted averages for the road segment length and the speed limit of the roadway were estimated using equation (3-1) and (3-2) to minimize subjectivity in digitization of the roadways:

$$Average\ Speed\ Limit\ (AVE\_SPEEDLIM) = \frac{\sum_{i=i}^{j} S_i n_i}{\sum_{i=1}^{j} N_i} \qquad \textit{(3-1)}$$

where:

S_i represent the total speed limit of road in a particular subcategory, e.g. 20 km/hr;

$n_i$is the total sample size of speed limit; and

N_i is the total number of speed classes that falls within a TAZ.

Similarly, average road segment length was calculated using the same approach (see Equation 3-2).

$$Average\ Segment\ Length\ (AVE\_SEGLEN) = \frac{\sum_{i=i}^{j} L_i n_i}{\sum_{i=1}^{j} N_i} \qquad \textit{(3-2)}$$

where:

L_i is the total length of road class within a joint – arterial, collector etc;

n_i is the total sample size of the road class; and

N_i is the total population size of road class within each class.

A land use database was extracted from the shape files obtained from open.regina.ca. This shape file was overlaid with Regina's TAZs to determine the proportion of each land use type within each zone. The summaries of the collisions, crimes, segments by road category, speed and functional class, intersections by number of legs, and proportion of each land use was merged with the TAZ file that contained the sociodemographic information.

In total, two separate databases containing aggregated information for Regina's TAZs were generated for macro-level modelling: the aggregated crime database and the aggregated collision database.

However, aggregation biases could not be completely overcome. One such issue was the problem of boundary crimes and collisions.

3.4 Boundary Crimes and Collision

Apart from the spatial heterogeneity/dependency problem that characterizes areal unit analysis, another problem peculiar to macro-level analysis is the treatment of incidences that occur at the boundary of two or more zones, i.e., boundary incidences assignment. TAZs are in most cases delineated using the existing boundary between zones. In some cases, the boundaries of two or more zones are defined by arterial or street networks which often result in collisions occurring at or near the boundary of zones. Cui et al., (2015) explained that a large proportion of road traffic collisions occur at the boundary and if underestimated or overestimated, the results of the analysis, especially in the case of road safety, could be biased.

While modelling aggregated zonal level events such as collisions and crimes involves assigning the events into zones, the issue of events that occur at or near a boundary of two TAZs remains problematic (Siddiqui and Abdel-Aty, 2012). Various methods of classifying boundary collisions have been explored, e.g., ignoring the boundary event, assigning events on boundaries to both areas (double counting shared events), using a weighting scheme, and using a boundary density ratio method such as that recently proposed by Cui et al., (2015).

Boundary assumptions could play an important role in zonal-level model estimation. The simplicity of the boundary assumption influences the sophistication of model built. Hence, it is important to evaluate alternative treatments that could be given to boundary collisions and crimes.

Ladron de Guevara et al. (2004) employed an approach which focuses on eliminating collisions that occur at the boundary of TAZs. This approach was use in the development of long-range collision prediction models for the city of Tucson in Arizona. They pointed out that due to differences in the level at which data were collected, two types of random error were generated. The first type relates to the treatment given to an arterial or street segment on the boundary of two zones. Such segments cannot be uniquely assigned to a zone. Ladron de Guevara et al. suggested that these segments should be removed as they accounted for only 5% of the entire road length within their study. The second type of random error relates to treatments given collisions information assigned to points. Collisions that occurred on boundary were simply removed and not considered because they produced random errors when attempts were made to allocate them into TAZs. This treatment of boundary collisions was used in research by Hadayeghi (2009) despite Ladron de Guevara et al. pointed out that eliminating boundary collisions introduces random error which was claimed to negligible.

The approach to removing boundary collisions ignores the fact that most TAZs are delineated by roads (usually arterials) and collisions on arterials are common due to high traffic volumes. It also ignores the fact that boundary collisions situations vary by city and that the case of Tucson, Arizona might apply to other cities. Neglecting such collisions may result in a larger proportion of collisions not being considered in the model development. The approach used by Ladron de Guevera et al. (2004) is, simple to use, but not logical for the development of aggregated collision prediction models. The assumption of negligible random error is broad and requires verification of the proportion of collisions that fall on the boundary to avoid a large proportion of collisions possibly being ignore. Hence its use may largely depend on the evaluation of the proportion of collisions that occurs on the boundary of TAZs. The assumption of zonal influences on boundary collisions was completely ignored by Ladron de Guevera et al.

In the study by Wang et al. (2012), a pre-weighted approach was highlighted and said to be proposed by Sun and Lovegroove (2010). This method was also used by Sun and Lovegroove (2013). It involves sharing collisions that occur at the boundary between the number of TAZs involved (sharing point features by corresponding TAZs by pre-rating them with a weight equal to the reciprocals of the related TAZs). This method of allocating boundary collisions is quite logical and simple to use, but it neglects the problem that traffic crosses several boundaries. Simply dividing the boundary collisions between the TAZs sharing a boundary road may not accurately account for boundary collisions or relate the collisions to the appropriate TAZ. This technique of pre-weighting might seem appropriate when information about the numbers of trips crossing the boundary of traffic zones is not available.

Wang et al., (2012) suggested that three steps should be evaluated when attempting to allocate boundary collisions to TAZs. When intersections are shared by different TAZs, Wang et al. suggested that the direction of travel and the configuration of accidents between vehicles could be used for allocating collisions into zones where an intersection is shared by more than one TAZ. When roadways are shared by TAZs, the side of the road where the collision occurs is used to allocate the collisions into zones. When information about the first two steps is unavailable, Wang et al. recommend the pre-rating method.

Of all the methods suggested, only the pre-weighted technique appears feasible. It is simple and does not completely violate the principle that neighbouring zones influence boundary collisions. The concept of examining collision configuration before assigning collisions to zones may be less feasible where details regarding travel direction, collision configuration and side of the road information where the collision occurred is not available.

Another approach suggested for treating boundary collisions involves development of separate models for interior and boundary collisions. In an attempt to study the nature of collisions at the boundary of zones, Siddiqui and Abdel-Aty (2012) developed pedestrian collisions prediction models using data from the counties of Hillsborough and Pinellas in Florida. Collisions that occurred within traffic zones were referred to as interior collisions and those that occurred at the boundaries as defined by a buffer distances were termed boundary collisions. Collisions that occurred within 100 meters of the TAZ border were termed boundary collisions. Separate models were developed for interior and boundary collisions using a hierarchical Bayesian method based on the roadway characteristics and socio-demographic characteristics of zones. The results of the models were compared with TAZ level models that did not consider boundary effects or the influence of neighbouring zones on pedestrian collisions.

Developing separate models for interior and boundary collisions was found to enhance the prediction of pedestrian collisions. Models that ignored boundary effects on pedestrian collisions were found to underperform. While this approach presents an opportunity for the treatment of boundary collisions, it could be considered only suitable for pedestrian collisions. This is because pedestrians do not travel long distance as evidence by the research of Sidiqui and Abdel-Aty (2012) as 70% of the collisions in the two counties occurred within the border. Also, roads have different widths, and the use of buffers to determine boundary collisions may not effectively represent the geometry of road features that delineate traffic zones. Similarly, defining a unique buffer when vehicle to vehicle collisions is studied by severity could be challenging.

Lee et al. (2014) attempted to address boundary issue problem by proposing the development of Traffic Safety Analysis Zones (TSAZs). TSAZs were created by combining TAZs that have homogenous numbers of collisions into new zones. This led to the creation of a reduced number of zones. The Brown-Forsythe test of homogeneity of variance was employed to check similarity in zones before aggregation. An optimal zone aggregation of 1:2 was recommended when regionalization is considered. It was highlighted that developing TSAZs showed potential in reducing the boundary issues of TAZs. While development of TSAZs led to a reduction in the number of boundary collisions, the majority of collisions (about 62%) still occurred on or near a boundary. This defeated the purpose of creating TSAZs to minimize boundary collision issues. The approach may, however, be effective for the change of support problem in area unit modelling. Lee et al. recommended that further research should be done on boundary collisions as regionalization does not eradicate the boundary collision problem in long range transportation planning.

Recently, Cui et al. (2015) proposed a three-step approach to boundary collision aggregation. The approach is known as the collision density ratio (CDR). In this method, collisions that occurred at the boundary of two or more TAZs were determined using histogram-based entropy. This determined the size of the boundary zones and collisions that occurred at a boundary. Probability density distributions for collisions within each zone were then evaluated and the CDR was used to aggregate the collisions into appropriate zones. The method proposed by Cui et al. was evaluated using the city of Edmonton as a case study to classify boundary collisions. The results obtained appeared promising as entropy-based histogram was found suitable for identifying boundary collisions with high accuracy. The authors claimed that the CDR method was more suitable than the weighted or one to one approach to boundary collision classifications and aggregations. While boundary collisions (otherwise referred to as the boundary effect in the literature) has been receiving wide attention, the authors recognized that roads in reality are not lines as represented in spatial analysis. In reality, the thin lines used to represent road have width, and where a description of the road's lanes where these collisions occur is not given, the weighted approach may be appropriate. The complexity involved in this approach may cause some limitation in its usage. It does not appropriately take into consideration the intra-zonal effect of zones on collisions. The assumption of the approach proposed by Cui et al. that the traffic conditions in a neighbourhood have an influence on boundary collisions is bogus as there is no appropriate justification for intra-zonal trips occurring between zones. In this case, simply dividing boundary collisions by the numbers of joints that shared a boundary is more appropriate.

While different approaches exist to dealing with boundary collisions, the choice depends on simplicity, the information available, and the nature of the collisions.

Most crimes do not occur at the boundary of TAZs, but where they do occur on a boundary, dividing the crimes between the of zones that share the boundary was considered suitable.

3.5 Applied Technique for the Allocation of Boundary Crimes and Collisions

Crime and collisions occurring at the boundary of areal units are common in any attempt to aggregate individual level information into areal units. This type of problem is not peculiar to TAZs. It is widely known to be occur when an attempt is made to aggregate information into diverse units. Figure 3-6 shows an example of collision incidences occurring at the boundary of two or more areal units. Different methods have been proposed for handling such incidences in transportation safety engineering. However, none is known to exist in crimes studies although boundary crime incidences can also occur.

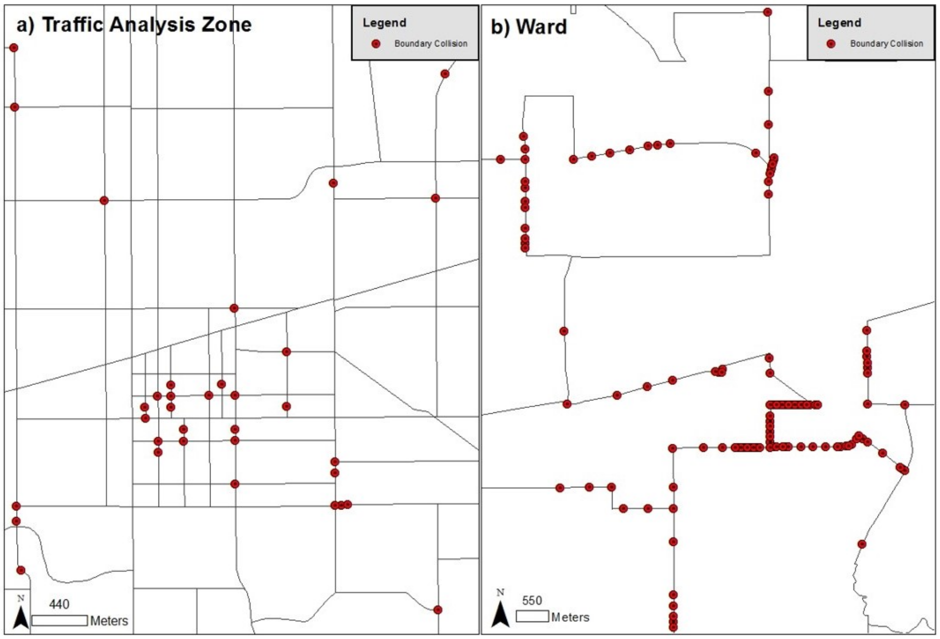


Figure 3-6: Collision Incidences on Boundary of Zones

In this research, the method of pre-weighting that simply divides incidences occurring at the boundary of two or more areal units by the number of zones that share the crimes or collisions was employed. The mathematical description of the approach used for aggregating boundary incidences is given as Equation (3-3):

$$Pre - rated = \frac{Total\ Number\ of\ Incidences\ at\ the\ Boundary}{Number\ of\ Zones\ that\ Shared\ the\ Boundary} \tag{3-3}$$

The incidences of interest in this research could be crimes or collisions. This approach was found useful in aggregating boundary events because it assumes that zones that share such boundary exert equal influence. However, the approach is not without bias. Consider two neighbouring zones that has one collision at their boundary. Using the pre-weighting technique assumes that 0.5 collisions or crimes will be added to the neighboring zone. Eventually, this will round up to one, implying that one collision each is added to these unit. Also, in cases where collisions occurred at the boundary of the city, this implies that potentially half of these collisions might not be accounted for. These are aggregation biases that characterize this method. Thus, the pre-weighting technique to boundary collisions does not completely overcome aggregation biases. Nonetheless, in the creation of databases for analysis that will be discussed, the pre-weighting method was extensively used in assigning boundary data.

Importantly in this research, in the collision database, where the pre-weighting method assigned collisions to areas with no exposure to traffic collisions, the collisions were reassigned equally to zones within the same proximity.

3.6 Research Approach and Modelling Technique

In this research, three modelling techniques were considered in the development of the crime and collision prediction models for hotspot identification. These models included the negative binomial regression method which is the state-of-the-art model for count data. The spatial variant of count data model considered in this research included the geographically weighted Poisson (GWPR) and negative binomial regression (GWNBR) techniques. The use of negative binomial regression in this analysis was to facilitate selection of appropriate variables that could be used to calibrate geographically weighted Poisson and negative binomial regression models. The Moran I test was employed to provide justification for the use of the spatial variant of these models. Predictor variables selected for the crimes categories and collisions by severity were tested for the presence of spatial dependency. If spatial dependence existed, there was adequate justification of the need for spatial models.

The flow process used in this research is shown on Figure 3-7.

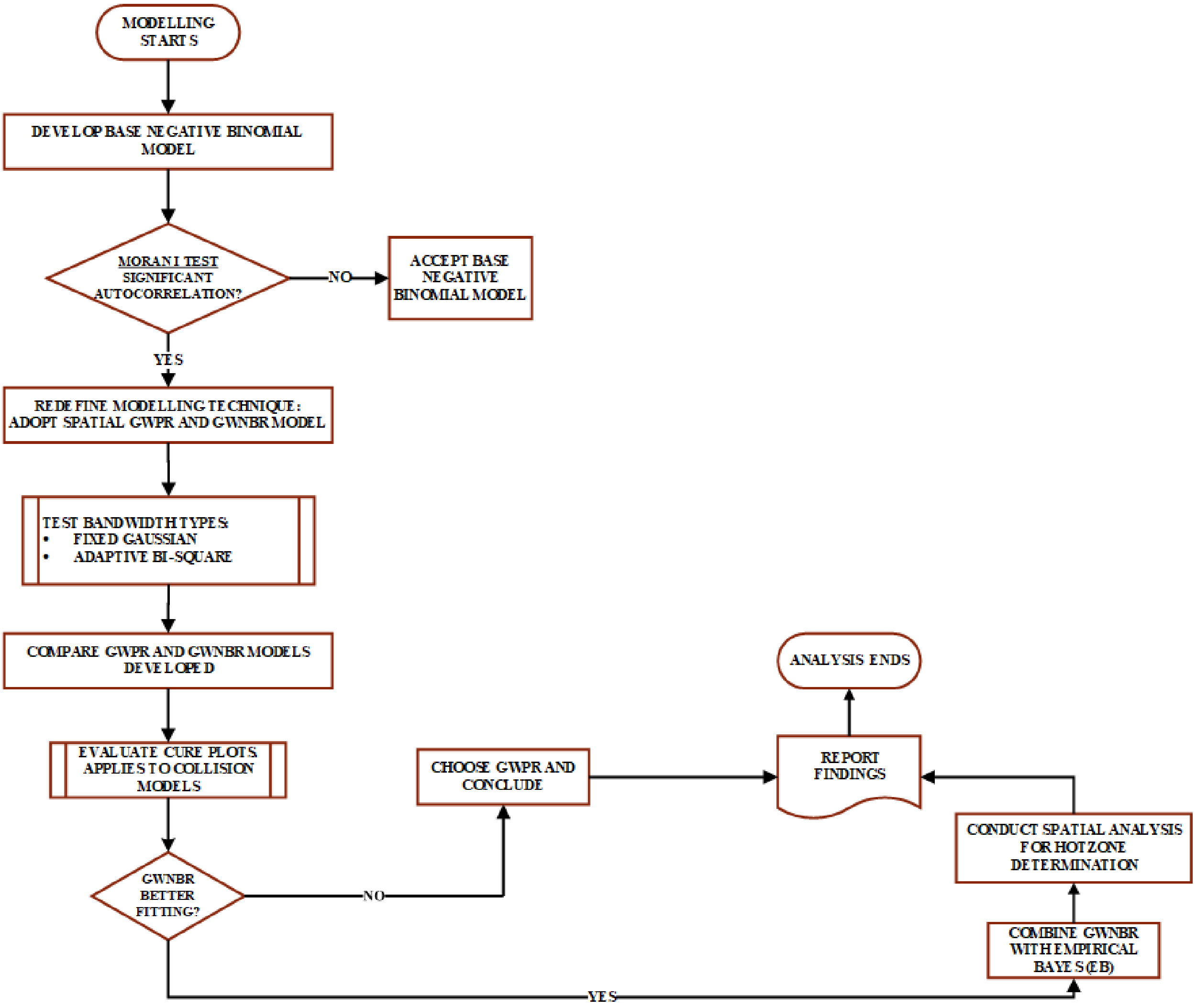

MODELLING STARTS
DEVELOP BASE NEGATIVE BINOMIAL MODEL
MORAN I TEST SIGNIFICANT AUTOCORRELATION?
NO
ACCEPT BASE NEGATIVE BINOMIAL MODEL
YES
REDEFINE MODELLING TECHNIQUE: ADOPT SPATIAL GWPR AND GWNBR MODEL
TEST BANDWIDTH TYPES:
• FIXED GAUSSIAN
• ADAPTIVE BI-SQUARE
COMPARE GWPR AND GWNBR MODELS DEVELOPED
EVALUATE CURE PLOTS APPLIES TO COLLISION MODELS
GWNBR BETTER FITTING?
NO
CHOOSE GWPR AND CONCLUDE
REPORT FINDINGS
ANALYSIS ENDS
CONDUCT SPATIAL ANALYSIS FOR HOT ZONE DETERMINATION
COMBINE GWNBR WITH EMPIRICAL BAYES (EB)
YES

Figure 3-7: Schematic Flow Diagram of Research Approach

This research compares the performance of geographically weighted regressions (the Poisson and the negative binomial types) considering two bandwidth choices: the fixed Gaussian bandwidth and the adaptive bi-square bandwidth. These bandwidths were compared to determine their influence on the performance of the models developed. Upon comparison of the models' goodness of fit and cumulative residual (CURE) plots (CURE plots were applied to collision models with known exposure (VKMT)), the Empirical Bayes (EB) technique was applied to enhance precision in hotspot identification and to correct for regression-to-the-mean bias. A description of measures of model goodness of fit available to this research for the selection of the best performing geographically weighted regression technique based on bandwidth choice is provided in section 3.7.

3.7 Measures of Goodness of Fit

Goodness of fit measures are used to evaluate how well a model predicts the observed data. Different tests can be used to measure the performance of any statistical models. Some tests use the log likelihood information of models. These tests include the Akaike Information Criterion, corrected Akaike Information Criterion, and the Bayesian Information criteria. Others determine the performance of models by considering the difference between the observed data and the predicted value. These tests include the Mean Square Error, Mean Square Prediction Error, Mean Absolute Deviation, and Mean Prediction Bias. Different researchers may use different tests to evaluate model performance.

It is important to evaluate all available measures to determine consistency in model performance. The main goodness of fit tests are discussed in sections 3.7.1 through 3.7.7.

3.7.1 Akaike Information Criteria (AIC)

This is used to find models that best explain the data with minimum free parameters. It penalizes models with a number of parameters (Burnham and Anderson, 2004; Young and Park, 2013). The equation for estimating AIC of a model is given as equation (3-4):

$$AIC = -2\, log\, L + 2(p) \qquad \textit{(3-4)}$$

L and p in Equation 3-4 represent the model log likelihood and the number of estimated parameters estimated in the model.

3.7.2 Corrected Akaike Information Criteria (AICc)

The corrected Akaike information criteria is similar to the initial Akaike information criteria except that it penalizes for sample size and the number of predictors used in model calibration. AIC_c adjusts for bias that might be inherent in model estimation due to small sample size (Burnham and Anderson, 2004). In most cases, AIC_c converges to AIC as the sample size approaches infinity (Akpa and Unuabonah, 2011). The formula for estimation of AIC_c is given in Equation (3-5):

$$AIC_c = AIC + \frac{2p^2+2p}{N-p-1} \qquad \textit{(3-5)}$$

The parameters in this equation are the same as Equation (3-4), but N is the sample size.

3.7.3 Bayesian Information Criteria (BIC)

Similar to Akaike Information criterion, BIC is used to evaluate model goodness of fit. BIC was proposed by Schwarz (1978). It is used for independently identically distributed data with the assumption of exponential distribution. However, unlike AIC, BIC uses a sample estimator based on transformation of posterior distribution associated with the models. Equation (3-6) gives the formula for estimating BIC of a model:

$$BIC = -2\, log\, L + 2\, log\, N\, p \qquad \textit{(3-6)}$$

The parameters of equation (3-5) and (3-6) are the same.

3.7.4 Mean Square Error (MSE)

The Mean Square Error is carried out on estimation data and measures how closely the fitted value represents the observed value: the smaller the mean square error, the closer the fit to the actual value. A model with small a MSE is preferred (Oh, et al., 2003; Young and Park, 2013). The MSE is estimated using Equation (3-7):

$$MSE = \frac{\sum_{i=1}^{N}(Y_i - \mu_i)^2}{N-p} \tag{3-7}$$

3.7.5 Mean Square Prediction Error (MSPE)

The Mean Square Prediction Error measures how closely the fitted value represents the observed value: the smaller the mean square error, the closer the fit to the actual value (Oh, et al., 2003; Young and Park, 2013). A model with a small MSPE is preferred. The test is carried out on test data. The MSPE is estimated using Equation (3-8):

$$MSPE = \frac{\sum_{i=1}^{N}(Y_i - \mu_i)^2}{N} \tag{3-8}$$

3.7.6 Mean Prediction Bias (MPB)

Mean prediction Bias (MPB) explores the magnitude and direction of model bias (Giuffre et al, 2014; Oh, et al., 2003; Oh et al., 2004; Young and Park, 2013). Positive MPB means that the model is over predicting while a negative value means that the model is under predicting. A value of zero is the most desirable but practically impossible. A model with an MPB value close to zero is preferred. The MPB is estimated using Equation (3-9):

$$MPB = \frac{\sum_{i=1}^{N}(Y_i - \mu_i)}{N} \tag{3-9}$$

3.7.7 Mean Absolute Deviation (MAD)

This measures how well the model predicts, i.e., it measures average model misspecification. A model with values close to zero is considered the best and is preferred (Oh, et al., 2003; Oh et al., 2004; Young and Park, 2013). The formula for estimation of the MPB is shown in Equation (3-10):

$$MAD = \frac{\sum_{i=1}^{N} |Y_i - \mu_i|}{N} \qquad \textit{(3-10)}$$

From equations (3-7) to (3-10), Y_i, μ_i and N are the observed count of crimes or collisions, the predicted value, and the sample size of that data respectively, and p is the number of parameters in the model.

3.7.8 Cumulative Residual (CURE) Plot

Another diagnostic measure for evaluating the performance of a model is the cumulative residual plot. This has been used especially in collision modelling owing to the availability of a common exposure variable that that can be used. This method was originally proposed for use in transportation safety studies by Hauer and Bamfor (1997). It measures how well a model predicts for a given range of exposure variable. It is a useful technique that can be employed to check and adjust the fit of a model. Basically, CURE plots evaluate the performance of a model using the residual (i.e., the difference between the observed and the predicted values) to assess a model fit. An example of a cumulative residual plot developed at a zonal level is given in Figure 3-8. VKMT within each zone is used to measure exposure.

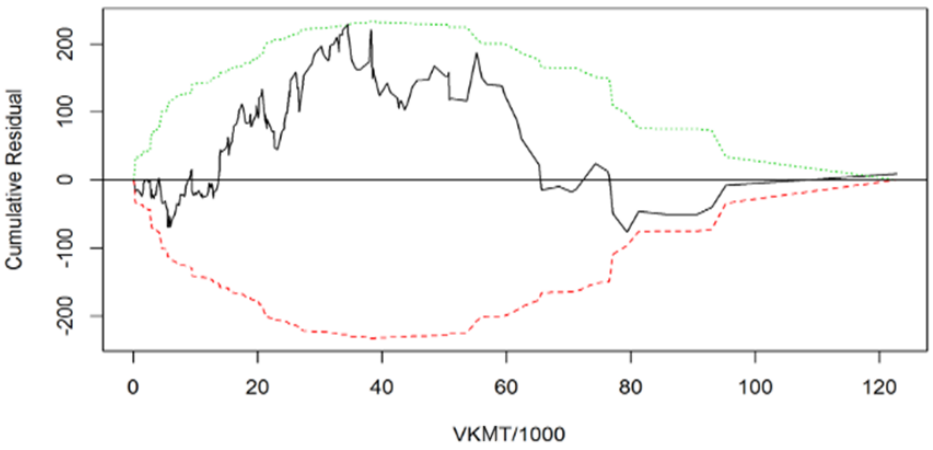


Figure 3-8: An Example of a CURE Plot for a Collision Model

(Takyi et al., 2018)

When the exposure variable (VKMT) was between 17 and 35, the observed value was larger than the predicted count collision data (see Figure 3-8). However, when the VKMT was above 35 and below 75, the model overestimated. The number of observed collisions was less than the number of predicted collisions. Beyond this, the model underestimated. A good CURE plot is expected to revolve around the zero coordinates and converge to zero. A model with a poor fit has the cumulative residual line entirely below or above the zero coordinates. In general, when the cumulative residual shows a tight band with the zero coordinate, the predicted values are much closer to the observed collisions. CURE plot has been found particularly useful in assessing the fit of models especially in collision studies. While CURE plots can be used to evaluate the fitting performance of models, there are some limitation to their use. CURE plots show the pattern of the cumulative residual and not the predictive value along a given range of exposure. Also, when two CURE plots are similar, selecting the one that better fits or represents the data could be challenging.

References

Akpa, O. M., and Unuabonah, E. I. (2011). Small-sample corrected Akaike information criterion: an appropriate statistical tool for ranking of adsorption isotherm models. Desalination, 272 (1-3), 20-26.

Blincoe, L., Seay, A., Zaloshnja, E., Miller, T., Romano, E., Luchter, S., and Spicer, R. (2002). The economic impact of motor vehicle crashes, 2000. DOT HS, 809, 446.

Boyce, J., Cotter, A., and Perreault, S. (2014). Police-reported crime statistics in Canada, 2013. Juristat: Canadian Centre for Justice Statistics, 1.

Burnham, K. P., and Anderson, D. R. (2003). Model selection and multimodel inference: a practical information-theoretic approach. Springer Science & Business Media.

Burnham, K. P., and Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. Sociological methods and research, 33(2), 261-304.

Cui, G., Wang, X., and Kwon, D. W., (2015). A framework of boundary collision data aggregation into neighborhoods. Accident Analysis and Prevention, 83, 1-17.

Giuffrè, O., Granà, A., Giuffrè, T., Marino, R., and Marino, S. (2014). Estimating the safety performance function for urban un-signalized four-legged one-way intersections in Palermo, Italy. Archives of Civil Engineering, vol. 60, No.1, pp. 41-54.

Hadayeghi, A., (2009). Use of advanced techniques to estimate zonal level safety planning models and examine their temporal transferability. Ph.D. Thesis. Department of Civil Engineering, University of Toronto.

Hauer, E., and Bamfo, J. (1997, November). Two tools for finding what function links the dependent variable to the explanatory variables. In proceedings of the ICTCT 1997 Conference, Lund, Sweden.

Highway Safety Manual (2010). American Association of State Highway and Transportation Officials (AASHTO), 1st Edition, Washington DC, ISBN: 9781560514770

Ladron de Guevara, F., Washington, S., and Oh, J. (2004). Forecasting crashes at the planning level: simultaneous negative binomial crash model applied in Tucson, Arizona. Transportation Research Record: Journal of the Transportation Research Board, (1897), 191-199.

Lee, J., Abdel-Aty, M., and Jiang, X. (2014). Development of zone system for macro-level traffic safety analysis. Journal of transport geography, 38, 13-21.

Oh, J., Lyon, C., Washington, S., Persaud, B., and Bared, J. (2003). Validation of FHWA crash models for rural intersections: Lessons learned. Transportation Research Record: Journal of the Transportation Research Board, (1840), 41-49.

Oh, J., Washington, S., and Choi, K. (2004). Development of accident prediction models for rural highway intersections. Transportation Research Record: Journal of the Transportation Research Board, (1897), 18-27.

Saskatchewan Government Insurance: SGI (2013). 2013 Saskatchewan traffic accident facts - Annual Report.

Schwarz, G., (1978). Estimating the dimension of a model. The Annals of Statistics, 6(2), 461-464.

Siddiqui, C., and Abdel-Aty, M. (2012). Nature of modeling boundary pedestrian crashes at zones. Transportation Research Record: Journal of the Transportation Research Board, (2299), 31-40.

Statistics Canada (2013). Police-reported crime statistics in Canada, 2013. https://www150.statcan.gc.ca/n1/pub/85-002-x/2014001/article/14040-eng.htm

Statistics Canada (2017a). Canada at a glance: population. https://www150.statcan.gc.ca/n1/pub/12-581-x/2017000/pop-eng.htm

Statistics Canada. (2017b). Focus on geography series, 2016 census. Statistics Canada Catalogue no. 98-404-X2016001. Ottawa, Ontario. Data products, 2016 Census.

Sun, J., and Lovegrove, G. (2013). Comparing the road safety of neighbourhood development patterns: traditional versus sustainable communities. Canadian Journal of Civil Engineering, 40(1), 35-45.

Takyi, E. A., Oluwajana, S. D., and Park, P. Y. (2018). Development of macro-level crime and collision prediction models to support data-driven approach to crime and traffic safety (DDACTS). Transportation Research Record, 0361198118777356.

The City of Regina (2018). About Regina. https://www.regina.ca/residents/about regina/index.html

Wang, X., Jin, Y., Abdel-Aty, M., Tremont, P., and Chen, X. (2012). Macrolevel model development for safety assessment of road network structures. Transportation Research Record: Journal of the Transportation Research Board, (2280), 100-109.

Young, J., and Park, P. Y. (2013). Benefits of small municipalities using jurisdiction-specific safety performance functions rather than the Highway Safety Manual's calibrated or uncalibrated safety performance functions. Canadian Journal of Civil Engineering, 40(6), 517-527.

CHAPTER 4:DEVELOPMENT OF MACRO-LEVEL COLLISION PREDICTION MODELS

4.1 Problem Statement

This chapter develops a macro-level collision prediction model. (See Chapter 5 for the development of the crime models.) Macro-level collision prediction models, also known as zonal-level or aggregated level models, are becoming increasingly popular among transportation safety professionals. Macro-level models are used to estimate the safety performance of individual zones in a study region. These models assist in screening the level of safety in traffic analysis zones (TAZs) and/or evaluating the introduction of safety countermeasures (e.g., adjusting speed limits) in the screened zones. The estimated number of collisions per zone obtained from a macro-level collision prediction model is often used as a safety performance measure. Past studies have tended to use generalized linear models, such as negative binomial (NB) regression models that use zonal-level input variables (e.g., socio-demographic data, traffic, and land use data) to estimate the safety performance of individual zones in a study region (Hadayeghi et al, 2003; Lovegroove and Sayed, 2006; Lovegroove et al., 2009; Pirdavani et al., 2012; Moeinaddini et al., 2015).

The use of zonal-level socio-demographic, traffic and/or land use data as input variables in a statistical model to estimate certain traffic parameters is not new to transportation engineers. Trip generation models, for example, have long used socio-demographic and/or land use data as input variables to estimate, for instance, the number of trips to and from zones as part of the well-established four-step travel forecasting model (Easa, 1993; Wilmot, 1995; Martin, and McGuckin., 1998; De Dios Ortuza and Willumsen, 2010; Kwigizile and Teng, 2009; Johnson et al., 2016). In any zonal model, heterogeneity within and across zones is a methodological challenge. This includes heterogeneity in the various input variables across zones and unobserved heterogeneity within zones.

Mannering et al. (2016) discussed the nature and characteristics of unobserved heterogeneity in detail and provided several examples that may introduce unobserved heterogeneity into a collision prediction model. They noted that although driver age is a common input variable (used, for example, to establish a statistical relationship between driver age and the occurrence and/or severity of collisions), driver age may be a serious source of unobserved heterogeneity. Driver age could have a very wide range of impact on individual drivers' driving behaviours. As a result, the researchers argued that a person's age may in reality be merely a proxy variable. Modellers may be using driver age to represent the impact of many unobserved variables such as an individual driver's physical characteristics or risk-taking behavior. It is not possible for modellers to collect such information from every driver. The unobserved heterogeneity in driver age will be exacerbated when age is aggregated (e.g., 16 to 24, 25 to 40, etc.). Mannering et al. (2016) reported that unobserved heterogeneity can introduce model misspecification and can also produce bias in the parameters estimated by the model. The model's inferences and/or predictions are then less reliable.

Unobserved heterogeneity can be a serious issue for a macro-level collision prediction model that uses TAZs as the spatial analysis unit especially if the socio-demographic and land use data associated with the TAZs are limited. Serious issues can arise with small variations in an input variable across the zones within a study region. For example, the number of intersections in a zone may be used as an input variable to show the statistical relationship between the number of intersections and the number of collisions in each zone. However, zones with a very similar (or even identical) number of intersections could show a very different number of collisions. The difference could be due to unobserved factors that affect the traffic and/or infrastructure condition of intersections in each zone differently for intersections for which the data were not collected. If these factors are not part of the model, they cannot be used to explain differences in the number of collisions in each zone. In the case of intersections, examples of unobserved factors that may not be included in the study data could include intersection configuration (e.g., skewed or perpendicular), intersection sight distance, the presence/absence of a heavy vehicle route through an intersection, and the presence/absence of left and/or right turn exclusive lanes at an intersection.

One particularly important and popular input variable for a macro-level collision prediction model is the vehicle kilometers travelled (VKMT) per zone. Unfortunately, this variable is also subject to unobserved heterogeneity due to spatial variation in traffic volume in a zone. For instance, a zone may be dominated by a few major arterials/freeways that carry most of the traffic. Another zone might have a similar VKMT, but no major arterials/freeways. In this case, the traffic will be more evenly distributed across the zone. The number and severity of collisions in the two zones may be very different although the VKMT for each zone is similar.

For various practical, financial, privacy, and/or political reasons, it is not possible to avoid unobserved heterogeneity in the input variables used in a macro-level collision prediction model. Mannering et al. (2016) suggest that modellers should work to develop advanced techniques that can help to reduce unobserved heterogeneity and avoid model misspecification in macro-level collision prediction models.

4.2 Methodological Challenges

Spatial Non-Stationary Parameters

Conventional regression methods assume that the model parameters (intercept and slope) remain constant or universal for entire zones in a study region (Brunsdon et al., 1996; Charlton and Fotheringham, 2009; Fotheringham et al., 1998; Fotheringham et al., 2003; Fotheringham and Oshan, 2016; Huang et al., 2012; Mennis, 2006), but it may be difficult to explain variation in the collision data of many zones in terms of the selected input variables and the constant parameters associated with the input variables. This means that the assumption of constant parameters in each zone must be relaxed if there is need to improve explanation of variation beyond the explanation provided by the input variables.

An advanced technique known as a random parameter model has become available relatively recently. This model allows variation in parameters. It allows parameters to vary (e.g., varying intercepts and/or varying slopes) across study observations (zones in the case of the study like this one) and helps to address the problem of unobserved heterogeneity in collision prediction modelling. Applications in road safety include Anastasopoulos et al (2012), Xu and Huang (2015), Chen and Tarko (2014), and Dong et al., (2014). The models are designed to take into account variation in collisions across zones not only in terms of variation in the input variables of each zone, but also in terms of the variation in parameters of each zone. This means that even in the case of zones with very similar (or even identical) values for the input variables, it is possible to consider the possibly different (i.e., random) parameters for different zones and estimate different numbers and/or severity of collisions for the different zones.

Although the random parameter model is a powerful tool that can take care of a certain level of unobserved heterogeneity in collision data, the model requires the modeller to pre-specify the distribution of the parameters (the distribution of the intercepts and/or slopes) to allow the parameters to vary across observations. Unfortunately, some studies based on simulation analysis have suggested that mis-specifying the distribution of the random parameters can yield severely biased estimates of all model parameters (Neuhaus et al., 1992; Neuhaus et al., 2013).

Spatial Dependency

Spatial dependency in the collision data is another challenge for random parameter models. Spatial dependence in collision data means that collisions that occur close to each other may be associated with factors that are spatially correlated with each other. At a micro-level, for example, traffic signal coordination at multiple intersections along a corridor may act as a common factor in collisions at the intersections with the coordinated traffic signals. Independently controlled intersections may show less spatial dependency. At a macro or zonal level, a downtown area may include multiple zones that are close to each other and have similar lane use characteristics. These downtown zones may generate a similar number/severity of collisions (per unit area) due to similarities in the roadway surface infrastructure (e.g., lack of shoulder lanes or absence of roadside parking space), traffic controls (e.g., one-way streets), travel behaviours (e.g., large number of commuters using public transit), and/or vehicle classification (e.g., similar proportion of freight transportation). Non-downtown zones are less likely to show such spatial dependency.

The possibility of spatial dependency in collision and other input variables led modellers to take spatial dependence in input variables across observations (zones in this study) into account. Yu (2010) used an advanced statistical modelling technique called geographically weighted regression (GWR). GWR is designed to take spatial dependency in input data into consideration.

GWR also incorporates the idea that a model's constant parameters are not suitable for explaining possibly large amounts of variation in collision data. An advantage of GWR over other methods, such as the random parameter model, lies in the way that GWR estimates parameters. Unlike random parameter models, GWR does not require each parameter's distribution to be specified. This is helpful as a parameter's distribution may vary from one study dataset to another and the form of the distribution is usually unknown to modellers.

GWR does not assume that the parameters vary randomly. It assumes that the parameters vary rather deterministically with the relative locations within a study area. A macro-level GWR collision prediction model assumes that the input variables for zones that are close to each other are more strongly related to each other than those for zones that are further away from each other. Parameter estimation in GWR is also less computationally demanding than in some random parameter models (Finley, 2011). A major weakness of GWR is the assumption that the response is normally distributed (Nakaya et al., 2005). This assumption is not acceptable for collision data (i.e., count data).

GWR has evolved into the geographically weighted Poisson regression (GWPR) model in which the error-term can be non-normally distributed. GWPR was first introduced by Nakaya et al. (2005). The model was developed based on the principle of generalized linear models described by Nelder and Wedderburn (1972). Unfortunately, the GWPR model, like a conventional Poisson regression model, cannot handle over-dispersed count data and this is a typical circumstance for many collision data (Hauer, 2001; Lord and Mannering, 2010). Interestingly, although GWPR is not ideal for handling over-dispersed collision data, Xu and Huang (2015) reported that the fitting capability (i.e., the predictive performance) of their GWPR model outperformed a random parameter model when estimating the number of collisions particularly in the case of collision data with spatial dependency.

Da Silva and Rodrigues (2014) introduced a geographically weighted negative binomial regression model (GWNBR) to improve handling of over-dispersed data. Their model can consider unobserved heterogeneity and spatial dependency simultaneously with varying parameters. The GWNBR has not yet been explored extensively as a tool for using zonal-level collision prediction to improve roadway safety.

This study explores the GWNBR. It focuses on the development of a set of macro-level collision prediction models using two different statistical modelling approaches: 1) GWPR and 2) GWNBR. The results from these models were compared using five years of collision and other zonal-level input data collected in Regina, Saskatchewan, Canada. Some methodological issues related to improving the fitting capability of the GWPR and GWNBR models were explored. For example, GWPR and GWNBR models construct different equations for zones falling within the bandwidth of each target zone. The method of choosing bandwidth may influence the parameters estimated (Faber and Paez, 2007; Guo et al., 2008; Paez et al., 2011). This chapter examines two different ways of considering bandwidth (fixed Gaussian and adaptive bi-square) for GWPR and GWNBR with the purpose of improving fitting accuracy.

4.3 Study Objectives

This chapter has two main objectives:

1. To develop macro-level collision prediction models based on two geographically weighted regression techniques (GWPR and GWNBR) to take into account the issue of spatial dependency in collision data; and

2. To evaluate the predictive performance of the GWPR and GWNBR models by considering, for instance, different ways of determining bandwidth in parameter estimation.

4.4. Study Data

The study used five years (2009 to 2013) of Regina collision, traffic volume, socio-demographic, and land use data to calibrate the macro-level collision prediction models. Five different databases from three different sources provided the data:

- Five years (2009 to 2013) of collision data (Microsoft Access format) were supplied by SGI;
- Five matching years of traffic volume data (GIS shapefile format) were supplied by the City of Regina;
- Roadway inventory data (GIS shapefile format) were supplied by the City of Regina. The data included the location of every intersection and roadway segment in the city (i.e., the roadway network), and details such as the functional classification and speed limit of every roadway segment; and
- Socio-demographic data (Microsoft Excel format) for each TAZ in Regina were supplied by the City of Regina. As the area of each TAZ was not included in this database, the area of each TAZ was estimated using the area information (GIS shapefile format) provided at open.regina.ca. This website is an online open data source platform supported by the City of Regina.

4.4.1 Collision Data

SGI records all collisions in Saskatchewan in a database known as the “Traffic Accident Information System” (TAIS). TAIS contain collision frequencies by severity and traffic control.

The dataset contains vehicle-to-vehicle collisions and distinguishes multiple vehicle collisions from single vehicle collisions, but multiple and single vehicle collisions were aggregated and used in the development of macro-level collision prediction models.

During the five-year study period, there were 26,642 collisions in Regina. Most (20,883 or 78.3%) were property damage only (PDO) collisions, and 5,759 (21.7%) were fatal or injury collisions (see Table 4-1).

4.4.2 Traffic Volume Data

The traffic volume data contained average annual daily traffic (AADT) data for the year 2013 for most of the roads in Regina, but some AADT data were missing. Local roadways in residential areas, in particular, tended to lack AADT information. Where necessary, the AADT were estimated (by calculating the average AADT for roadways with the same classification for which AADT data were available) and used to produce the VKMT for each TAZ. VKMT was used as the main exposure variable (called TAZ_VKMT) for collision modelling (see Table 4-1).

4.4.3 Roadway Inventory Data

As mentioned, the roadway inventory data included the functional classification and speed limit of each roadway segment. This information was used to generate the input variables in Table 4-1 for each TAZ in the city. Examples of input variables include the total length of roadways by functional classification (e.g., arterial length; ARTERIAL_LEN) and the length of roadway with each speed limit (e.g., total roadway length with 80 km/h posted speed limit; TOT_SEGLEN_80) per TAZ. Variables such as the number of intersections with a certain number of legs (e.g., number of intersections with 4 legs; NO_4LEGS_INT) per TAZ were also estimated by counting the intersections using ArcGIS. In total, 40 roadway variables were generated and used in the collision modelling. The 40 roadway variables are listed in Table 4-1.

4.4.4 Socio-Demographic Data

The TAZ socio-demographic and land use variables are shown in Table 4-1. The socio-demographic variables include total population (TOT_POP), a breakdown by age (e.g., those aged 65+; POP_65plus) and measures of population density. The land use variables include the percentage of each TAZ's area used by office space (OFFICE_SPACE) and by industry (INDUSTRIAL_AREA), etc. These data were obtained from the City of Regina. In total, 9 socio-demographic variables were used in the collision modelling.

In summary, 66 input variables were included in the study (one traffic volume variable, 40 road inventory variables, 9 socio-demographic variables, and 16 land use variables).

Table 4-1: Descriptive Statistics of Regina Collisions and Other Input Data

Input Variable	Description	Mean	Std. Dev	Min.	Max.
Collision Data					
TOT_PDO_CRASHES	Total Numbers of Property Damage Only Collision	85.59	68.07	1.00	326.00
TOT_FI_CRASHES	Total Number of Fatal Injury Collision	23.60	21.60	0.00	103.00
TOT_CRASHES	Total Number of Collision	109.18	87.57	1.00	425.00
Traffic Volume					
TAZ_VKMT	Vehicle kilometers travelled per Traffic Analysis Zone	25034.04	21770.69	321.90	122800.88
Road Variable					
PARKING_COST	Parking Cost	0.70	1.81	0.00	5.95
ARTERIAL_LEN	Arterial Length (km)	0.57	0.57	0.00	2.62

Table 4-1: Descriptive Statistics of Regina Collisions and Other Input Data (Cont'd)

Input Variable	Description	Mean	Std. Dev	Min.	Max.
Road Variable (Cont'd)					
COLLECTOR_LEN	Collector Length (km)	0.66	0.77	0.00	3.45
DRIVEWAY_LEN	Driveway Length (km)	0.00	0.01	0.00	0.17
EXPRESSWAY_LEN	Expressway Length (km)	0.09	0.27	0.00	2.03
GRAVEL_ROAD_LEN	Gravel Road Length (km)	0.11	0.46	0.00	4.88
HIGHWAY_LEN	Highway Length (km)	0.04	0.21	0.00	1.82
LOCAL_ROAD_LEN	Local Road Length (km)	2.52	2.41	0.00	9.64
PRIVATE_ROAD_LEN	Private Road Length (km)	0.25	1.03	0.00	12.73
RAMP_LEN	Ramp Length (km)	0.10	0.28	0.00	1.83
ROW_LEN	Right of Way Length (km)	0.00	0.03	0.00	0.43
TOT_SEGLEN	Total Segment Length (km)	4.34	3.34	0.09	16.35
ARTERIAL_DEN	Arterial Density (Length/Area)	2.00	3.39	0.00	40.76
COLLECTOR_DEN	Collector Density (Length/Area)	1.80	2.25	0.00	13.29
DRIVEWAY_DEN	Driveway Density (Length/Area)	0.00	0.01	0.00	0.23
EXPRESSWAY_DEN	Expressway Density (Length/Area)	0.23	0.79	0.00	7.46
GRAVEL_ROAD_DEN	Gravel Road Density (Length/Area)	0.10	0.38	0.00	3.13
HIGHWAY_DEN	Highway Density (Length/Area)	0.06	0.32	0.00	2.79
LOCAL_ROAD_DEN	Local Road Density (Length/Area)	5.76	4.04	0.00	15.27
PRIVATE_ROAD_DEN	Private Density (Length/Area)	0.37	1.13	0.00	11.07
RAMP_DEN	Ramp Density (Length/Area)	0.18	0.55	0.00	3.28
ROW_DEN	Right of way Density (Length/Area)	0.02	0.22	0.00	3.28
TOT_SEGLEN_DEN	Total Segment Length Density (Length/Area)	10.53	5.20	0.82	40.76
NO_SEG_PER_TAZ	Numbers of Segments per Traffic Analysis Zone	28.36	24.26	1.00	134.00
AVE_SPDLIM	Weighted Average Speed Limit (km/hr.)	84.27	56.65	40.00	536.17
AVE_SEGLEN	Weighted Average Segment Length (km)	2.23	1.87	0.08	11.36
TOT_SEGLEN_20	Roadway length with Posted Speed Limit 20 km/hr.	0.00	0.01	0.00	0.20
TOT_SEGLEN_30	Roadway length with Posted Speed Limit 30 km/hr.	0.00	0.05	0.00	0.52
TOT_SEGLEN_40	Roadway length with Posted Speed Limit 40 km/hr.	0.68	0.83	0.00	3.84

Table 4-1: Descriptive Statistics of Regina Collisions and Other Input Data (Cont’d)

Input Variable	Description	Mean	Std. Dev	Min.	Max.
Road Variable (Cont'd)					
TOT_SEGLEN_50	Roadway length with Posted Speed Limit 50 km/hr.	3.33	2.70	0.00	14.41
TOT_SEGLEN_60	Roadway length with Posted Speed Limit 60 km/hr.	0.03	0.16	0.00	1.70
TOT_SEGLEN_70	Roadway length with Posted Speed Limit 70 km/hr.	0.12	0.38	0.00	3.28
TOT_SEGLEN_80	Roadway length with Posted Speed Limit 80 km/hr.	0.09	0.37	0.00	3.68
TOT_SEGLEN_100	Roadway length with Posted Speed Limit 100 km/hr.	0.07	0.26	0.00	1.82
NO_3LEGS_INT	Numbers of Intersections with 3 Legs	9.66	10.10	0.00	66.00
NO_4LEGS_INT	Numbers of Intersections with 4 Legs	5.59	6.49	0.00	36.00
NO_5LEGS_INT	Numbers of Intersections with 5 Legs	0.02	0.13	0.00	1.00
TOTAL_INT	Total Number of Intersections	15.25	13.60	0.00	79.00
SEG_DEN	Segment Density (Number of Segments/Area)	72.90	43.60	0.83	244.12
INT_DEN	Intersection Density (Number of Intersections/Area)	38.13	24.40	0.00	129.20
Socio-demography					
POP_0to17	Proportion of Persons Age 0 to 17	0.17	0.09	0.00	0.29
POP_18to24	Proportion of Persons Age 18 to 24	0.09	0.04	0.00	0.33
POP_25to44	Proportion of Persons Age 25 to 44	0.25	0.10	0.00	0.40
POP_45to64	Proportion of Persons Age 45 to 64	0.23	0.09	0.00	0.34
POP_65plus	Proportion of Persons Age 65 plus	0.16	0.13	0.00	0.55
TOT_POP	Total Population	808.31	798.35	0.00	3011.00
POP_DENSITY	Population Density (persons/km^2)	2121.49	1673.68	0.00	10552.61
NO_GRDSCH	Proportion of Persons Enrolled in Graduate School	0.10	0.33	0.00	3.90
NO_PSSTUD	Proportion of Persons Enrolled in Post-Secondary School	0.43	5.09	0.00	73.31

Table 4-1: Descriptive Statistics of Regina Collisions and Other Input Data (Cont'd)

Input Variable	Description	Mean	Std. Dev	Min.	Max.
Land Use					
OFFICE_SPACE	Proportion of Office Floor Space Area	0.06	0.21	0.00	1.90
RETAIL_SPACE	Proportion of Retail Floor Space Area	0.04	0.09	0.00	0.78
INDUSTRY_SPACE	Proportion of Industry Floor Space Area	0.01	0.03	0.00	0.20
HOSPT_SPACE	Proportion of Hospital Floor Space Area	0.00	0.02	0.00	0.24
NO_LU_PER_TAZ	Number of Land Uses Per Traffic Analysis Zone	4.22	1.65	1.00	8.00
AIRPORT_AREA	Proportion of Airport Area per Traffic Analysis Zone	0.00	0.05	0.00	0.85
COMMERCIAL_AREA	Proportion of Commercial Area per Traffic Analysis Zone	0.22	0.35	0.00	1.00
INDUSTRIAL_AREA	Proportion of Industrial Area per Traffic Analysis Zone	0.09	0.26	0.00	1.00
INSTITUTIONAL_AREA	Proportion of Institutional Area per Traffic Analysis Zone	0.03	0.08	0.00	0.88
OPENSPACE_RECREATION _AREA	Proportion of Open space/Recreational Area per Traffic Analysis Zone	0.14	0.23	0.00	1.00
RAILWAY_AREA	Proportion of Railway Area per Traffic Analysis Zone	0.02	0.09	0.00	1.00
RESIDENTIAL_HD_AREA	Proportion of Residential High-Density Area per Traffic Analysis Zone	0.04	0.08	0.00	0.62
RESIDENTIAL_LD_AREA	Proportion of Residential Low-Density Area per Traffic Analysis Zone	0.34	0.33	0.00	0.97
RESIDENTIAL_MD_AREA	Proportion of Residential Medium Density Area per Traffic Analysis Zone	0.04	0.12	0.00	0.91
URBAN_HOLDING_AREA	Proportion of Urban Holding Area per Traffic Analysis Zone	0.07	0.22	0.00	1.00
TAZ_AREA	Traffic Analysis Zone Area (km^2)	0.51	0.56	0.01	6.04

4.5 Methodology

Collision prediction models for three different levels of severity, i.e. total, property damage only (PDO) and fatal-injury (FI), were initially developed using a geographically weighted regression technique. The main tool was SAS University Edition. Fifteen different collision prediction models were developed and evaluated. As PDO collisions accounted for 78.3% of the collisions, the input variables and estimated parameters for the total and PDO models were very similar. As a result, the study described in this chapter concentrates on the 10 collision prediction models for FI and PDO collisions only. The models for total collisions are not discussed. The discussion of methodology used in this study is presented in the following five sections:

- Section 4.5.1 presents the result of Moran's I local indicator which was used to measure spatial dependence. Dummy models (one FI and one PDO) using conventional NB regression were also developed. The results of the dummy models are reported in Appendix, but Moran's I local indicator confirmed the possibility of spatial dependency in important input variables in the study data;
- Section 4.5.2 presents the model specifications for the GWPR models used in this chapter;
- Section 4.5.3 presents the model specifications for the GWNBR models used in this chapter;
- Section 4.5.4 presents the form of the two distinct types of bandwidth (fixed Gaussian and adaptive bi-square bandwidth) investigated on the predictive performance of the collision prediction models developed. A single optimal bandwidth (fixed Gaussian bandwidth) for an entire study region has usually been used in geographically weighted regression modelling (Rhee et al., 2016). Varying bandwidth (adaptive bi-square bandwidth) has also been considered, for example, by using small bandwidths for high density areas (e.g., downtown)

and large bandwidths for low density areas (e.g., non-downtown). Some studies have reported that adaptive bi-square bandwidth approaches have advantages over fixed Gaussian bandwidth approaches in geographically weighted regression applications (Hadayeghi et al., 2010; Li et al., 2013; Pirdavani et al., 2014; Yao et al., 2015; Shariat-Mohaymany, et al., 2015; Amoh-Gyimah et al., 2017); and

- Section 4.5.5 presents the multiple goodness-of-fit (GOF) tests including the cumulative residual (CURE) plots employed in the study. The GOF tests and CURE plots provided a basis for the rigorous evaluation of the collision prediction models developed.

4.5.1 Moran's I Local Indicator for Testing Spatial Dependency

Moran's I local indicator was used to examine whether the study data had spatial dependence issues that required special consideration. Moran's I local indicator is designed to measure how observations of selected variables in a zone are similar or dissimilar to the observations of the same variables in surrounding zones (Pirdavani et al., 2014). If the observations for nearby zones are similar, the observations may not be statistically independent. Lack of statistical independence violates the assumption required by many conventional statistical models, including conventional negative binomial model, that the data are independent. The presence of dependence in a study data renders the use of conventional negative binomial models invalid and supports the use of GWPR or GWNBR models.

Moran's I local indicator ($\mathrm{I_i}$) for spatial dependency is given as Equation (4-1). The results are reported in Section 4.7.

$$\boldsymbol{I_i} = \frac{\sum_{j=1, j \neq i}^{n} (x_i - \bar{x})(x_j - \bar{x})}{s_x^2 \sum_{j=1}^{n} w_{ij}} \qquad \textit{(4-1)}$$

In Equation (4-1), x_i and x_j represent the input variables for zones i and j, $\bar{x}$ is the mean of the input variables. n is the sample size (i.e., the numbers of traffic analysis zones in this study), w_{ij} is the geographic distance weighting scheme between the centroid of zones i and j, and s is the standard deviation of x.

4.5.2 Geographically Weighted Poisson Regression (GWPR)

Consider a Poisson model parameterized in terms of the collision rate ${}^{\mu_i}/_{t_j}$ given in Equation (4-2) where t_j is the exposure time. This is often considered as an offset variable. u_i is the predicted mean.

$$u_i = exp\left(t_j \sum_{i=1}^{k} \beta_i x_{ij}\right) \tag{4-2}$$

By allowing spatially varying parameters across $u - v$ space, Equation (4-2) becomes:

$$y_i \sim Poisson\left[t_j exp\left(\sum_{i=1}^{k} \beta_i\left(u_j, v_j\right) x_{ij}\right)\right] \tag{4-3}$$

where:

β_i is the parameter of the input variable x_i for $i = 1, \dots \dots, K$;

y_i is the i^{th} dependent variable for zone i; and

$\left(u_j, v_j\right)$ represents the location coordinates for zone *i*.

The most popularly used functional form of collision prediction models discussed in the literature is shown in Equation (4-4) (Hadegeyi et al., 2010; Li et al., 2013; Pirdavani et al., 2014):

$$ln(Y) = ln\left(\beta_0\left(u_j, v_j\right)\right) + \beta_1\left(u_j, v_j\right)\ ln(VKMT) + \cdots\cdots + \beta_k\left(u_j, v_j\right) X_k + \varepsilon \tag{4-4}$$

The parameters $\beta(u_j, v_j)$ of this model are allowed to vary spatially and are described in Equation (4-5):

$$\beta(u_j, v_j) = \begin{bmatrix} \beta_0(u_1, v_1) & \beta_1(u_1, v_1) & \beta_2(u_1, v_1) & \cdots & \beta_k(u_1, v_1) \\ \beta_0(u_2, v_2) & \beta_1(u_2, v_2) & \beta_2(u_2, v_2) & \cdots & \beta_k(u_2, v_2) \\ \cdots & \cdots & \cdots & & \cdots \ \cdots \\ \beta_0(u_n, v_n) & \beta_1(u_n, v_n) & \beta_2(u_n, v_n) & \cdots & \beta_k(u_n, v_n) \end{bmatrix} \tag{4-5}$$

For any given regression zone i referenced by geographic coordinate (u_i, v_i), the numerical solution for parameters $\beta(u_i, v_i)$ can be obtained using Equation (4-6):

$$\beta(u_i, v_i) = (X^T W(u_i, v_i) X)^{-1} X^T W(u_i, v_i) Y \tag{4-6}$$

where:

Y is the matrix of the dependent variable;

X is the $n \times k$ matrix of input variables; and

W is $n \times n$ the diagonal weighting matrix (obtained using bandwidth) for zone i, given as Equations (4-7) and (4-8):

$$X = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{i2} & \cdots & x_{2k} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} \tag{4-7}$$

$$W(u_i, v_i) = \begin{pmatrix} w_{i1} & 0 & \cdots & 0 \\ 0 & w_{i2} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & w_{in} \end{pmatrix} \tag{4-8}$$

$W(u_i, v_i)$ is derived from the bandwidth weighting function

4.5.3 Geographically Weighted Negative Binomial Regression (GWNBR)

As mentioned earlier, Da Silva and Rodriguez (2014) introduced the GWNBR model. The model can be viewed as an extension of GWPR and is expected to provide more robust results than the GWPR model for some collision datasets that show over-dispersion. The GWNBR model relaxes the assumptions of equal variance and equal mean that limit the Poisson model. The GWNBR model also allows the parameters of the models to vary spatially.

The general form of the negative binomial model is given in Equation (4-9). This equation is also parameterized in terms of the t_j (the offset variable) and u_i (the predicted mean):

$$u_i \sim NB\left[t_j\, exp\left(\sum_{i=1}^{k} \beta_i x_{ij}\right), \alpha\right] \qquad \textit{(4-9)}$$

where α in Equation (4-9) is the over-dispersion. The other parameters are the same as those in Equation (4-2). By allowing the parameters α and β_i of the model described in Equation (4-9) to vary locally, the functional form of the GWNBR model is given as Equation (4-10). Equation (4-10) is an extension of Equation (4-9):

$$y_i \sim NB\left[t_j\, exp\left(\sum_{i=1}^{k} \beta_i(u_j, v_j) x_{ij}\right), \alpha(u_j, v_j)\right] \qquad \textit{(4-10)}$$

The parameters of the negative binomial model described in Equation (4-9) are obtained iteratively using a combination of Newton Raphson (NR) and Iterative Reweighted Least Squares (IRLS). However, to allow for estimation of the locally varying parameter $\beta(u, v)$ of the negative binomial model, the NR and IRLS algorithms require modification. Da Silva and Rodriguez (2014) provide a full description of the algorithm modifications.

The log-likelihood of the GWNBR model is described in Equation (4-11) for a given regression zone i near j:

$$L\big(\beta(u_i,v_i)|x_{ij},y_i,\alpha_i\big) =$$

$$\sum_{i=1}^{n}\left\{\begin{matrix} y_j\, log\left[\alpha_i\mu_i\big(\beta(u_i,v_i)\big)\right] - \left[y_i + {}^{1}\!/_{\alpha_i}\right]log\left[1+\alpha_i\mu_i\big(\beta(u_i,v_i)\big)\right] \\ +\, log\left[\gamma\big(y_i + {}^{1}\!/_{\alpha_i}\big) - log\left[\gamma\big({}^{1}\!/_{\alpha_i}\big)\right] - log\left[\gamma(y_i+1)\right]\right] \end{matrix}\right\} w\big(d_{ij}\big) \qquad (4\text{-}11)$$

where $i = 1, \dots, n$, and $\mu_i\big(\beta(u_i,v_i)\big)$ is the predicted mean at zone i considering the parameter at zone j. This parameter is given as Equation (4-12):

$$\mu_i\big(\beta(u_i,v_i)\big) = t_j exp\big(\textstyle\sum_{i=1}^{k}\beta_i(u_i,v_i)x_{ij}\big) \qquad (4\text{-}12)$$

where $w\big(d_{ij}\big)$ in Equation (4-12) is the geographic weight for observation j at regression point i. $w\big(d_{ij}\big)$ depends on the distance between i and j.

The parameters of the GWNBR model can be obtained by maximizing the log-likelihood function given in Equation (4-12). In this study, the over-dispersion parameter was constrained by making it a constant value (i.e., constant α). Da Silva and Rodriguez (2014) pointed out that estimating constant α may generate a biased estimate for the over-dispersion parameter. Da Silva and Rodriguez (2014) also stated that making the over-dispersion parameter a constant value allows the calculation of the effective number of parameters of the model and the approach can still handle an overdispersed dataset (e.g., a typical collision data) better than can a GWPR model which considers α to be equal to zero. The GWNBR model with a constant dispersion parameter is described as Equation (4-13):

$$y_i{\sim}NB\left[t_j\, exp\big(\textstyle\sum_{i=1}^{k}\beta_i\big(u_j,v_j\big)x_{ij}\big)\,,\alpha\right] \qquad (4\text{-}13)$$

4.5.4 Bandwidth Type and Method of Optimal Selection

The equations for fixed Gaussian bandwidth and adaptive bi-square bandwidth are given in Equations (4-14) and (4-15) respectively.

$$w_{ij} = exp\left\{-\frac{1}{2}\left({d_{ij}}/{b}\right)^2\right\} \qquad \textit{(4-14)}$$

where w_{ij} represents the geographic weight of zone j when calibrating models for zone $i.$ In the case of fixed Gaussian bandwidth, d_{ij} is the distance between the j^{th} and i^{th} nearest neighbor zones, and b is the bandwidth.

$$w_{ij} = \begin{cases} \left[1-\left({d_{ij}}/{b}\right)^2\right]^2 & \textit{if } j \textit{ is one of the } N^{th} \textit{ nearest neighbor of } i \textit{ otherwise,} \\ 0 & \end{cases} \qquad \textit{(4-15)}$$

Cross-validation (CV) error was used to determine the optimal bandwidth for each type of bandwidth (Fotheringham et al., 2003; Pirdivani et al., 2014). The bandwidth that minimizes the cross-validation error was selected. Equation (4-16) shows the equation for the cross-validation error:

$$CV = \sum_{i=1}^{n}[y_i - \hat{y}_{\neq i}(b)]^2 \qquad \textit{(4-16)}$$

where:

CV is the cross-validation error;

$\hat{y}_{\neq i}(b)$ is the adjusted value to point i omitting the observation of i; and

n is the numbers of zones.

$\hat{y}_{\neq i}(b)$ is invariably the fitted value of y_i when the i^{th} zone is omitted during the calibration process.

4.5.5 Goodness-of-Fit (GOF) Tests

The selection of a good statistical model requires careful consideration of GOF tests. For example, a more complex model usually provides better GOF results, but may also result in over-fitting concerns. This study used seven GOF tests. The tests selected are probably the most popularly used tests (Schwarz, 1978; Washington, et al., 2005; Washington et al., 2010). These seven GOF tests are AIC, AIC_C, BIC, MSE, MSPE, MPB and MAD. They are described in the method section of chapter 3.

AIC, AIC_C and BIC rely on the likelihood function of a statistical model. These tests are shown in Equations (3-4) to (3-6). In these equations, LogL is the log likelihood function estimated from the models developed, p is the numbers of parameters in the model, and N is the sample size. As a statistical model adds parameters, the value of the likelihood function is usually increased, but the addition of the extra parameters may result in over-fitting. The AIC, AIC_C and BIC tests attempt to minimize this problem by adding a kind of penalty term to the likelihood function as extra parameters are added to a model. The BIC test adds the biggest penalty, and the AIC test adds the smallest. As a result, the BIC values are bigger than the AIC_C values which are bigger than the AIC values. A model with low AIC, AIC_C and/or BIC values is considered to have good GOF results.

The other four GOF tests are described in Equations (3-7) to (3-10). These four tests use the size of the residuals (the difference between the predicted (μ_i) and observed number of collisions (y_i) for zone i) to measure a model's predictive performance. The MPB test (Equation

(3-9)) measures both the magnitude and direction of model bias. A positive MPB indicates over prediction and a negative MPB underprediction. The MAD test (Equation (3-10)) also measures how well a model is predicting. Models with low values or close to zero on the MSE/MSPE/MPB/MAD tests are considered to have good GOF results and are preferred.

Seven tests were applied because no single GOF test can determine the best model to use. Different tests favour different models. Previous studies have noted this inconsistency in GOF test results and have emphasized the importance of using GOF tests carefully when determining the best fitting model (Koppius, 2009; Lord and Park, 2008; Shmueli, 2010; Young and Park, 2013). Hauer (2015) argued for using cumulative residual (CURE) plots rather than GOF tests when evaluating the fitting performance of a collision prediction model. Hence, a set of CURE plots was used in this study as one of the main tools for evaluating the collision models developed.

4.6 Results of Analysis

The two dummy models (one FI and one PDO) developed using a conventional negative binomial (NB) technique (see Appendix A for details of these models) were used to select the input variables that were statistically significant at the 90% confidence level. Table 4-2 lists the variables selected.

The dummy FI collision prediction model identified six statistically significant input variables: LOG_TAZ_VKMT, ARTERIAL_LEN, NO_LU_PER_TAZ, COMMERCIAL_AREA, INDUSTRIAL_AREA, and RESIDENTIAL_MD_AREA. The dummy PDO collision prediction model identified eight statistically significant input variables: LOG_TAZ_VKMT, INT_DEN, NO_3LEGS_INT, NO_LU_PER_TAZ, AVE_SEGLEN, COMMERCIAL_AREA, TOT_SEGLEN_DEN, and LOCAL_ROAD_DEN. As three variables

(LOG_TAZ_VKMT, NO_LU_PER_TAZ and COMMERCIAL_AREA) were statistically significant for both FI and PDO collisions, the total number of statistically significant input variables was 11.

Arterial road length (ARTERIAL_LEN) and industrial areas (INDUSTRIAL_AREA) were statistically significant variables for FI collisions, but not for PDO collisions. This might be expected as the higher speeds associated with arterials compared with most roadways are known to be a factor that contributes to FI collisions. Intersection density (INT_DEN) was a statistically significant variable for PDO collisions, but not for FI collisions. This might also be expected as zones with a higher intersection density will have more intersection turning movements and a higher number of conflicts than found in other zones.

4.7 Spatial Dependency

Moran's I Local Indicator was used to investigate spatial dependence for the 11 statistically significant input variables. The null hypothesis was no spatial dependency. The null hypothesis was rejected at the 95% confidence level for all 11 variables (Table 4-2). As this result indicated the presence of spatial dependence, thus GWPR/GWNBR model was preferable to a conventional Poisson and/or negative binomial regression model.

Table 4-2: Local Indicator for Spatial Dependence Statistics for Variables

Variables	Moran I_i	Z	Pr > |Z|
Dependent Variables			
TOT_FI_CRASHES	0.0266	10.3300	<.0001
TOT_PDO_CRASHES	0.0278	10.7590	<.0001
Independent Variables			
LOG_TAZ_VKMT	0.0650	23.2920	<.0001
INT_DEN	0.0574	20.7300	<.0001
NO_3LEGS_INT	0.0609	21.8800	<.0001

Table 4-2: Local Indicator for Spatial Dependence Statistics for Variables (Cont'd)

Variables	Moran I_i	Z	Pr > |Z|
NO_LU_PER_TAZ	0.0544	19.7034	<.0001
AVE_SEGLEN	0.0408	15.1400	<.0001
COMMERCIAL _AREA	0.0987	34.6400	<.0001
TOT_SEGLEN_DEN	0.0437	16.1000	<.0001
LOCAL_ROAD_DEN	0.0283	10.9180	<.0001
ARTERIAL_LEN	0.0170	7.1000	<.0001
INDUSTRIAL_AREA	0.0632	22.6600	<.0001
RESIDENTIAL_MD_AREA	0.0253	9.9100	<.0001

4.8 Model Parameters Estimated with Macro-Level Collision Prediction Models

The same 11 input variables identified as statistically significant in the dummy models were used to estimate the model parameters for the GWPR and GWNBR models. The parameter values vary from zone to zone and vary with the approach adopted to handle bandwidth (fixed Gaussian vs. adaptive bi-square). Section 4.8.1 discusses the results obtained from the GWPR models (fixed Gaussian and adaptive bi-square approaches), and Section 4.8.2 discusses the results obtained from the GWNBR models (fixed Gaussian and adaptive bi-square approaches). As mentioned in Section 4.5.4, the cross-validation method was used to determine optimal values for the fixed and adaptive bandwidth methods.

4.8.1 Geographically Weighted Poisson Regression (GWPR) Models

Table 4-3a summarizes the local parameters for the GWPR models using fixed Gaussian bandwidth, and Table 4-3b summarizes the local parameters for the GWPR models using adaptive bi-square bandwidth. The variation (range) of the parameters obtained for each input variable was examined and any change of sign was noted. It was clear that:

- In the case of variation in parameter values, there were similarities in the distribution of parameters estimated using fixed Gaussian and adaptive bi-square bandwidth; and
- In the case of sign reversal, models developed using a fixed Gaussian bandwidth approach showed sign reversal for four of the six input variables for FI collisions and six of the eight input variables for PDO collisions. Models developed using adaptive bi-square bandwidth approach also showed sign reversals, but not necessarily for the same input variables. The PDO model showed sign reversal for four of the eight variables (rather than six of eight). In general, the model using adaptive bi-square bandwidth produced a somewhat tighter range of parameter distributions for each variable than did the model using fixed Gaussian bandwidth.

LOG_VKMT provides an example of the wide range in parameter values estimated by the fixed Gaussian bandwidth approach. The estimated parameter values for LOG_VKMT ranged from 0.2704 to 1.6366 for FI collisions, and from 0.4119 to 1.0521 for PDO collisions (Table 4-3a). With adaptive bi-square bandwidth, the range varied from 0.2433 to 1.2936 for FI collisions, and from 0.4263 to 0.8795 for PDO collisions (Table 4-3b).

INDUSTRIAL_AREA provides an example of sign reversal when using a fixed Gaussian bandwidth approach for FI collisions. Most zones recorded an increase in FI collisions in industrial areas (positive sign), but INDUSTRIAL_AREA could also decrease the number of FI collisions (see value of -13.5151 in Table 4-3a).

ARTERIAL_LEN provides another example of sign reversal obtained when using a fixed Gaussian bandwidth approach for FI collisions. In most zones, longer arterials were associated with a higher estimate of the number of FI collisions, but longer arterials were also associated with a lower estimate of the number of FI collisions.

The variation and sign reversal found in the parameters estimated by the GWPR models developed using fixed Gaussian and adaptive bi-square bandwidth may have implications for the predictive performance of the models (see Section 4.9 for details). However, the variation and sign reversal found in the parameters makes the results difficult to interpret. It was suspected that sign reversal is partly due to unobserved heterogeneity within zones in the study dataset. In the case of arterials, for example, unobserved heterogeneity could be related to the presence, absence or type of median barriers, the level of street lighting, and/or the intensity of law enforcement along the arterials in each zone. These factors (and many others) were not included in the study data. As the wide range in parameters values appears to indicate the presence of variations that could not be explained by the input variables selected, the six variables selected for the FI models and the eight variables selected for the PDO models may be insufficient to explain the amount of variation in collisions across zones.

The GWPR models developed using an adaptive bi-square bandwidth (Table 4-3b) showed slightly less variation in the parameters estimated for the input variables and showed slightly fewer sign reversals. Although parameters for the two bandwidth choices were found to be similar in the models, it is difficult to use the distributions of the parameters reported in Table 4-3 to determine the models' prediction capability. (GOF tests used to gain additional insights into the models' prediction capabilities are discussed in Section 4.9.) The parameter variations in the GWPR models suggest that the models are effective in capturing unobserved heterogeneity in the dataset used in this study regardless of the bandwidth method employed.

Table 4-3a: Parameters Estimated by Geographically Weighted Poisson Model using Fixed Gaussian Bandwidth

Model	Parameter	Fixed Gaussian				
		Min	1st Quart	Median	3rd Quart	Max
Fatal-Injury	Intercept	-15.1092	-7.5625	-3.8236	-2.9248	-1.7386
	LOG_TAZ_VKMT	0.2704	0.3915	0.4798	0.8343	1.6366
	ARTERIAL_LEN	-0.4120	0.0320	0.1511	0.2461	0.8488
	NO_LU_PER_TAZ	-0.0660	0.0519	0.0915	0.1246	0.2681
	COMMERCIAL_AREA	0.1801	0.5378	0.6981	0.9872	3.0045
	INDUSTRIAL_AREA	-13.5151	0.0869	0.2171	0.4045	1.1605
	RESIDENTIAL_MD_AREA	-10.7846	0.3346	0.4417	0.7720	2.7014
	Over-dispersion (α)	Not Available (N/A)				
Property Damage Only	Intercept	-8.7215	-4.2452	-2.8076	-2.4381	-1.9921
	LOG_TAZ_VKMT	0.4119	0.5281	0.5800	0.6916	1.0521
	INT_DEN	-0.0023	0.0058	0.0091	0.0149	0.0576
	NO_3LEGS_INT	-0.0431	-0.0287	-0.0228	-0.0112	0.0186
	NO_LU_PER_TAZ	-0.1513	0.0032	0.0453	0.0743	0.2114
	AVE_SEGLEN	-0.2538	0.0445	0.0562	0.0764	0.2203
	COMMERCIAL_AREA	0.5313	0.7176	0.7860	0.9548	2.3993
	TOT_SEGLEN_DEN	-0.3169	-0.1243	-0.1110	-0.0707	0.0339
	LOCAL_ROAD_DEN	-0.0080	0.0741	0.0963	0.1017	0.1833
	Over-dispersion (α)	Not Available (N/A)				

Table 4-3b: Parameters Estimated by Geographically Weighted Poisson Model using Adaptive Bi-square Bandwidth

Model	Parameter	Adaptive Bi-square				
		Min	1st Quart	Median	3rd Quart	Max
Fatal-Injury	Intercept	-11.9193	-7.3430	-3.5369	-2.6767	-1.1560
	LOG_TAZ_VKMT	0.2433	0.3562	0.4592	0.8096	1.2936
	ARTERIAL_LEN	-0.3236	0.0359	0.1766	0.2795	0.6612
	NO_LU_PER_TAZ	-0.0020	0.0439	0.0826	0.1301	0.2232
	COMMERCIAL_AREA	0.1359	0.3981	0.7352	1.0408	2.3024
	INDUSTRIAL_AREA	-15.5083	-0.0868	0.1898	0.4299	5.0149
	RESIDENTIAL_MD_AREA	-5.3108	0.1895	0.3855	0.7619	2.0112
	Over-dispersion (α)	Not Available (N/A)				
Property Damage Only	Intercept	-6.7132	-3.7950	-2.8427	-2.4985	-1.8243
	LOG_TAZ_VKMT	0.4263	0.5193	0.5837	0.6436	0.8795
	INT_DEN	-0.0011	0.0068	0.0091	0.0143	0.0318
	NO_3LEGS_INT	-0.0356	-0.0285	-0.0226	-0.0100	0.0099
	NO_LU_PER_TAZ	-0.0634	0.0054	0.0423	0.0685	0.1622
	AVE_SEGLEN	-0.0289	0.0446	0.0548	0.0813	0.1032
	COMMERCIAL_AREA	0.4261	0.7261	0.8000	0.9782	1.3049
	TOT_SEGLEN_DEN	-0.2038	-0.1298	-0.1185	-0.0814	-0.0183
	LOCAL_ROAD_DEN	0.0573	0.0874	0.0991	0.1039	0.1469
	Over-dispersion (α)	Not Available (N/A)				

4.8.2 Geographically Weighted Negative Binomial Regression (GWNBR) Models

A set of GWNBR models was also developed to compare the parameter estimates and predictive performance of the GWNBR and GWPR models. The same input variables were used.

Tables 4-4a and 4-4b summarize the parameters estimated using the GWNBR models. The distribution of parameters was very similar to the GWPR models' parameter distribution. The variation in the parameters estimated using a fixed Gaussian bandwidth was similar to the parameters estimated using the adaptive bi-square bandwidth. Sign reversal occurred with both bandwidth approaches (as in the case of the GWPR models). The parameter variations in the GWNBR models again suggest that the models are effective in capturing unobserved heterogeneity in the dataset used regardless of the bandwidth method employed.

Table 4-4a: Estimated Parameters for Geographically Weighted Negative Binomial Model using Fixed Gaussian Bandwidth

Model	Parameters	Fixed Gaussian				
		Min	1st Quart	Median	3rd Quart	Max
Fatal-Injury	Intercept	-14.0117	-6.7964	-3.6755	-2.3729	-1.8128
	LOG_TAZ_VKMT	0.2628	0.3113	0.4457	0.7668	1.5297
	ARTERIAL_LEN	-0.4291	0.0140	0.1674	0.3018	0.7151
	NO_LU_PER_TAZ	-0.0785	0.0605	0.1176	0.1680	0.3080
	COMMERCIAL_AREA	0.2491	0.5855	0.7236	1.0422	4.7046
	INDUSTRIAL_AREA	-8.3916	0.1879	0.3130	0.5242	1.4987
	RESIDENTIAL_MD_AREA	-8.5707	0.4657	0.5608	0.9994	3.1586
	Over-dispersion (α)			1.9897		
Property Damaged Only	Intercept	-7.8181	-4.5430	-2.9918	-2.5776	-2.3091
	LOG_TAZ_VKMT	0.4985	0.5351	0.5669	0.6708	1.0112
	INT_DEN	-0.0107	0.0051	0.0072	0.0100	0.0671
	NO_3LEGS_INT	-0.0416	-0.0254	-0.0195	-0.0122	0.0269
	NO_LU_PER_TAZ	-0.0818	0.0274	0.0575	0.0854	0.2411
	AVE_SEGLEN	-0.1089	0.0371	0.0575	0.0741	0.1985
	COMMERCIAL_AREA	0.7159	0.8500	1.0071	1.3009	2.1249
	TOT_SEGLEN_DEN	-0.3021	-0.1040	-0.0950	-0.0671	0.1118
	LOCAL_ROAD_DEN	-0.0384	0.0862	0.0926	0.1114	0.1918
	Over-dispersion (α)			2.5747		

Table 4-4b: Parameters Estimated by Geographically Weighted Negative Binomial Model using Adaptive Bi-square Bandwidth

Model	Parameters	Adaptive Bi-square				
		Min	1st Quart	Median	3rd Quart	Max
Fatal-Injury	Intercept	-12.1038	-7.5803	-3.6198	-2.0872	-0.9719
	LOG_TAZ_VKMT	0.1970	0.2742	0.4473	0.8265	1.3393
	ARTERIAL_LEN	-0.5026	-0.0150	0.1982	0.3874	0.6391
	NO_LU_PER_TAZ	-0.0235	0.0544	0.1019	0.1806	0.2524
	COMMERCIAL_AREA	0.1566	0.4860	0.7422	1.1170	2.9724
	INDUSTRIAL_AREA	-15.9431	-0.0158	0.3049	0.5820	2.9117
	RESIDENTIAL_MD_AREA	-2.0136	0.3498	0.5254	1.0143	4.5854
	Over-dispersion (α)			1.9897		
Property Damaged Only	Intercept	-5.7848	-3.9386	-2.9697	-2.5174	-2.0378
	LOG_TAZ_VKMT	0.4924	0.5329	0.5663	0.6299	0.8270
	INT_DEN	0.0007	0.0056	0.0070	0.0083	0.0244
	NO_3LEGS_INT	-0.0351	-0.0262	-0.0196	-0.0120	0.0068
	NO_LU_PER_TAZ	-0.0078	0.0281	0.0545	0.0900	0.1480
	AVE_SEGLEN	-0.0673	0.0398	0.0595	0.0734	0.1072
	COMMERCIAL_AREA	0.5806	0.8580	1.0007	1.2093	1.8372
	TOT_SEGLEN_DEN	-0.2059	-0.1065	-0.0977	-0.0789	-0.0418
	LOCAL_ROAD_DEN	0.0695	0.0844	0.0968	0.1185	0.1902
	Over-dispersion (α)			2.5747		

4.9 Model Selection

Table 4-5 shows the results of the seven GOF tests for the GWPR and GWNBR models for FI and PDO collisions, and for the fixed Gaussian and adaptive bi-square bandwidth approaches. The log likelihood provided an eighth assessment of the models. The bold font in the Table highlights the best fitting model as assessed by each test.

Table 4-5: Goodness of Fit Test Results

Severity	Model	Method	Log Likelihood	AIC	AICC	BIC	MSE	MSPE	MPB	MAD
Fatal-Injury	GWPR	Fixed Gaussian	-1155.27	2402.61	2424.60	2563.61	5.77	5.60	-0.03	**1.65**
		Adaptive bi-square	-1195.11	2475.87	2494.63	2625.64	**5.50**	**5.35**	-0.24	1.71
	GWNBR	Fixed Gaussian	**-889.82**	**1857.81**	**1873.18**	**1994.51**	6.46	6.28	**0.00**	1.72
		Adaptive-bi-square	-897.19	1877.07	1894.44	2021.68	5.79	5.63	-0.22	1.73
Property Damage Only	GWPR	Fixed Gaussian	-2245.65	4585.80	4609.09	4751.04	**47.82**	**46.06**	**-0.07**	**5.06**
		Adaptive bi-square	-2641.94	5353.00	5364.80	5473.87	55.51	53.47	-0.61	5.63
	GWNBR	Fixed Gaussian	**-1184.09**	**2446.95**	**2462.58**	2584.71	58.84	56.67	0.10	5.52
		Adaptive bi-square	-1210.01	2474.78	2482.00	**2570.56**	63.89	61.53	-0.65	5.91

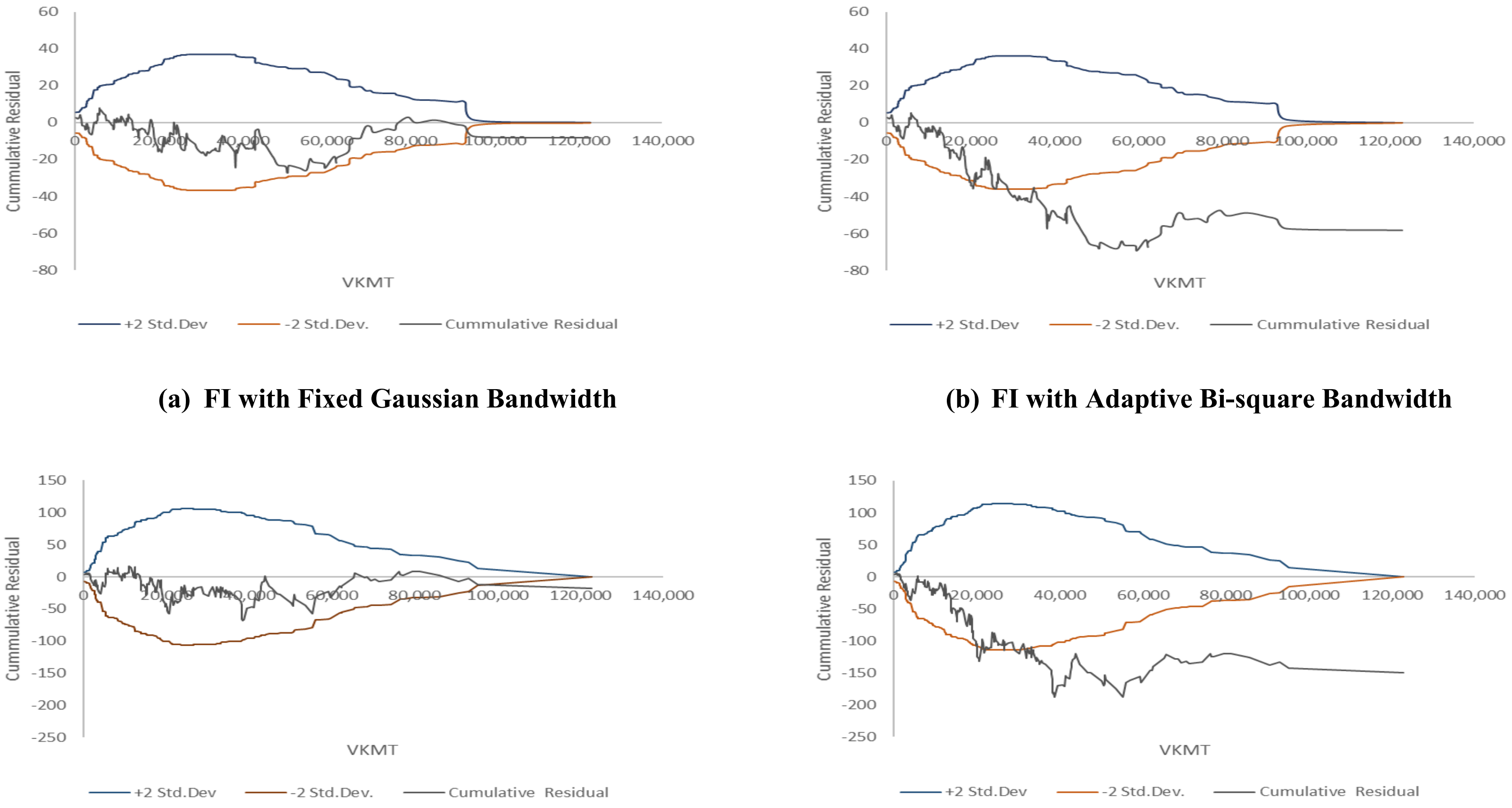
The results did not conclusively favour the GWPR or the GWNBR in terms of fitting performance:

- In the case of FI collision prediction, the GWPR models achieved three best scores compared with five for the GWNBR model;
- In the case of PDO collision prediction, the GWPR models achieved four best scores as did the GWNBR models;
- In the case of FI collision prediction, fixed Gaussian bandwidth achieved six best scores and adaptive bi-square bandwidth achieved only two; and
- In the case of PDO collision prediction, fixed Gaussian bandwidth achieved seven best scores and adaptive bi-square bandwidth approach achieved only one.

As mentioned in Section 4.5.5, similar inconsistency in GOF test results has been reported in other studies (Lord and Park, 2008; Young and Park, 2013). As a result, an additional GOF test, the CURE plot was introduced (see Figures 4-1 and 4-2 on the next pages).

Figures 4-1 and 4-2 show the CURE plots for the GWPR and GWNBR models. Both plots show the results for LOG_VKMT for FI and PDO collisions for fixed Gaussian and adaptive bi-square bandwidth. A good CURE plot should move around the zero coordinate of the plot and should tend to converge at zero. The cumulative residuals (black lines) for the fixed Gaussian bandwidth plots were much closer to zero than were the adaptive bi-square bandwidth plots. The ± 2 standard deviation error band (blue and red lines) were also closer to zero (tighter) for the fixed Gaussian bandwidth models.

The results of the CURE plot analyses clearly suggest that the best models for predicting collisions in the study area i.e. the City of Regina, were the GWPR and GWNBR models with fixed Gaussian bandwidth. The performance gap between the GWPR and GWNBR models was very small. When the GOF results for the fixed Gaussian and adaptive bi-square bandwidth GWPR and GWNBR models were compared, it was found that the GWNBR models with fixed Gaussian bandwidth had the best GOF results. This finding applies to FI and PDO collisions. Thus, GWNBR models with fixed Gaussian bandwidth were selected as the best models for predicting FI and PDO collisions. The choice was based on the results of the GOF tests and CURE plots.



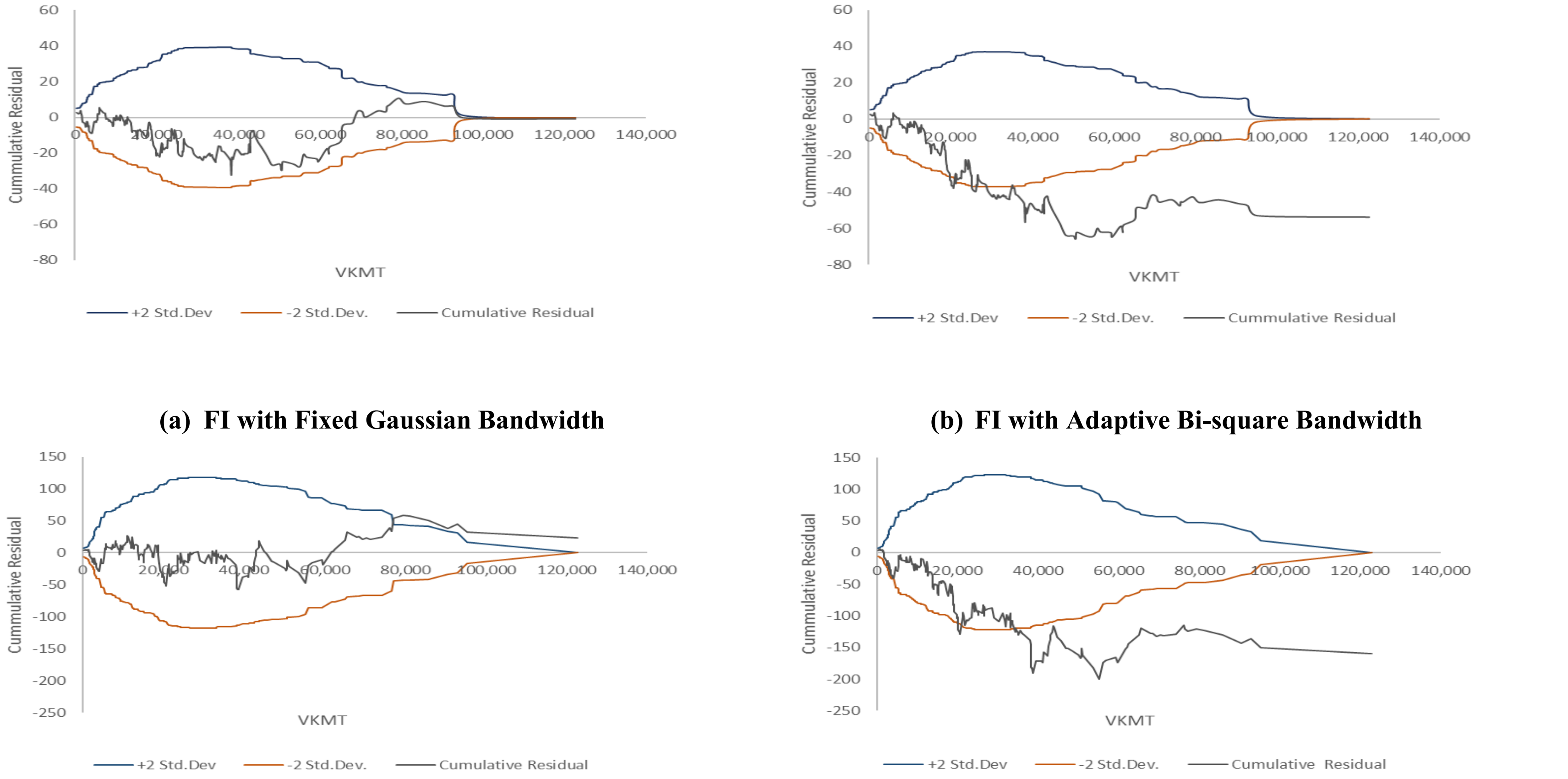
(a) FI with Fixed Gaussian Bandwidth

(b) FI with Adaptive Bi-square Bandwidth

(c) PDO with Fixed Gaussian Bandwidth

(d) PDO with Adaptive Bi-square Bandwidth

Figure 4-1: CURE Plots for GWPR for FI (a and b) and PDO (c and d) Collisions Showing Fixed Gaussian and Adaptive Bi-square Bandwidth Result



(a) FI with Fixed Gaussian Bandwidth

(b) FI with Adaptive Bi-square Bandwidth

(c) PDO with Fixed Gaussian Bandwidth

(d) PDO with Adaptive Bi-square Bandwidth

Figure 4-2: CURE Plots for GWNBR for FI (a and b) and PDO (c and d) Collisions Showing Fixed Gaussian and Adaptive Bi-square Bandwidth Result

4.10 Summary and Recommendations for Future Work

This study used a dataset for Regina, Saskatchewan, Canada to develop and compare two different types of geographically weighted regression model: GWPR and GWNBR. It initially considered 66 input variables. Using a set of conventional NB models, it was found that six of the 66 variables were statistically significant in FI collision prediction, and eight were statistically significant in PDO collision prediction, giving a total of 11 different variables (all significant at a confidence interval not less than 80%). No socio-demographic input variables were statistically significant. The statistically significant variables included traffic volume, road inventory variables (e.g., arterial length, intersection density), and land use variables (e.g., industrial area and commercial area).

Moran's I local indicator was used to check for spatial dependency in the variables. As all the selected variables showed spatial dependence, advanced models (GWPR and GWNBR) were used to handle this issue.

The study evaluated the impact of two different types of bandwidth (fixed Gaussian and adaptive bi-square) on the predictive performance of the GWPR and GWNBR collision prediction models. It used cross-validation error to select the optimal bandwidth for each approach. The results of the study showed that the type of bandwidth could significantly affect the predictive performance of the collision prediction models. Fixed Gaussian bandwidth appeared to offer improved predictive performance over adaptive bi-square bandwidths for all the GWPR and GWNBR models developed to predict the number of zonal levels FI and PDO collisions. The variation in the parameters of the models developed using fixed Gaussian bandwidth had a wider range than those developed using adaptive bi-square bandwidth. It was noticed that this wider range in the parameters helps to explain unobserved heterogeneity within each zone and helps to improve the predictive performance of the collision prediction models.

The results of the seven most popularly used GOF tests did not favour either the GWPR or the GWNBR models as the results of the GOF tests were not consistent. The CURE plots provided additional insights and helped us select the better performing model between GWPR and GWNBR. The GWNBR models were preferable to the GWPR models for explaining collision variation across zones in the study area considered although the performance gap between the two models was not large.

The smaller cumulative residuals, tighter error band, and greater convergence to zero of the fixed Gaussian bandwidth results compared to the adaptive bi-square bandwidth results suggested that the fixed Gaussian bandwidth models were preferable to the adaptive bi-square bandwidth models. The observations regarding the fixed Gaussian and adaptive bi-square bandwidth approaches cannot, however, be generalized for other collision datasets as the findings may be unique to this study's data. This issue requires additional investigation. Rather than rely on the experience of previous studies, it may be advisable to use the characteristics of the particular dataset to select the most appropriate bandwidth.

Future work should research a better way to select the optimal bandwidth for use in a GWPR or GWNBR model. The cross-validation method used to select the bandwidth used in the study produced similar results in terms of the individual parameter values, but suggested a different predictive performance for the different models. Future work should also investigate a better way to evaluate the predictive performance of the collision models.

Inconsistent results from the various GOF tests were also observed. CURE plots provide richer information than a single numeric GOF test result and appear to offer a potentially better technique.

A third area in need of research is the over-dispersion issue. As it is known that conventional NB models have an advantage over conventional Poisson models when handling an over-dispersed dataset, GWPR and GWNBR models should be compared to increase our understanding of how they handle over-dispersion.

Lastly, the processing time required to develop the GWPR and GWNBR models quantitatively was not analyzed. This issue was outside the scope of this study and it is understood that the processing time varies with the computer used, but it was noticed that a substantial amount of time (between 30 to 40 minutes) was needed to obtain calibrated parameters for each GWPR and GWNBR model developed. Improving optimization techniques for searching the parameters of GWPR and GWNBR models may help reduce the processing time and may encourage the application of GWPR and GWNBR models over conventional NB models wherever appropriate and feasible.

References

Amoh-Gyimah, R., Saberi, M., and Sarvi, M. (2017). The effect of variations in spatial units on unobserved heterogeneity in macroscopic crash models. Analytic Methods in Accident Research, 13, 28-51.

Anastasopoulos, P. C., Mannering, F. L., Shankar, V. N., and Haddock, J. E., (2012). A study of factors affecting highway accident rates using the random-parameters tobit model. Accident Analysis &and Prevention, 45, 628-633.

Brunsdon, C., Fotheringham, A.S., and Charlton M.E., (1996). Geographically weighted regression: A method for exploring spatial non-stationary. Geographical Analysis, 28(4), 281-298.

Charlton M., and Fotheringham A.S., (2009). Geographically weighted regression. White Paper. National Centre for Geocomputation, National University of Ireland Maynooth, Maynooth, Co Kildare, Ireland

Chen, E., and Tarko, A. P., (2014). Modeling safety of highway work zones with random parameters and random effects models. Analytic Methods in Accident Research, 1, 86-95.

Da Silva, A. R., and Rodrigues, T. C. V., (2014). Geographically weighted negative binomial regression—incorporating overdispersion. Statistics and Computing, 24(5), 769-783.

De Dios Ortuzar, J., and Willumsen, L. G., (2011). Modelling transport. New Jersey: Wiley. ISBN: 978-0-470-76039-0

Dong, C., Clarke, D. B., Yan, X., Khattak, A., and Huang, B. (2014). Multivariate random-parameters zero-inflated negative binomial regression model: An application to estimate crash frequencies at intersections. Accident Analysis & and Prevention, 70, 320-329.

Easa, S. M. (1993). Urban trip distribution in practice. I: Conventional analysis. Journal of Transportation Engineering, 119(6), 793-815.

Farber, S., and Páez, A., (2007). A systematic investigation of cross-validation in GWR model estimation: empirical analysis and Monte Carlo simulations. Journal of Geographical Systems, 9(4), 371-396.

Finley, A. O., (2011). Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence. Methods in Ecology and Evolution, 2(2), 143-154.

Fotheringham, A. S., Brunsdon, C., and Charlton, M., (2003). Geographically weighted regression: the analysis of spatially varying relationships. John Wiley and Sons.

Fotheringham, A. S., Charlton, M. E., and Brunsdon, C., (1998). Geographically weighted regression: a natural evolution of the expansion method for spatial data analysis. Environment and Planning A, 30(11), 1905-1927.

Fotheringham, A. S., and Oshan, T. M. (2016). Geographically weighted regression and multicollinearity: dispelling the myth. Journal of Geographical Systems, 18(4), 303-329.

Guo, L., Ma, Z., and Zhang, L. (2008). Comparison of bandwidth selection in application of geographically weighted regression: a case study. Canadian Journal of Forest Research, 38(9), 2526-2534.

Hadayeghi, A., Shalaby, A., and Persaud, B., (2003). Macrolevel accident prediction models for evaluating safety of urban transportation systems. Transportation Research Record: Journal of the Transportation Research Board, (1840), 87-95.

Hadayeghi, A., Shalaby, A. S., and Persaud, B. N. (2010). Development of planning level transportation safety tools using geographically weighted Poisson regression. Accident Analysis and Prevention, 42(2), 676-688.

Hauer, E., (2015). The art of regression modeling in road Safety, Springer.

Hauer, E., (2001). Overdispersion in modelling accidents on road sections and in empirical Bayes estimation. Accident Analysis and Prevention, 33(6), 799-808.

Huang B., Wu B., and Barry M., (2012). Geographically and temporally weighted regression for modeling Spatio-temporal variation in house prices. International Journal of Geographical Information Science, 24(3), 383–401

Johnson, E., Turochy, R. E., and LaMondia, J. J. (2016). Trip generation of student-oriented housing developments. Journal of Urban Planning and Development, 04016029.

Kwigizile, V., Teng, H., (2009). Comparison of methods for defining geographical connectivity for variables of trip generation models. Journal of Transportation Engineering, 135(7), 454-466.

Levine, N., Kim, K. E., and Nitz., L. H., (1995). Spatial analysis of Honolulu motor vehicle crashes: I. spatial patterns. Accident Analysis and Prevention, 27(5), 663-674.

Li Z., Wand W., Liu P., Bigham J.M., and Ragland D.R., (2013). Using geographical weighted Poisson regression for county level crash modeling. Safety Science 58, 89-97.

Lord, D., and Park, P. Y. J., (2008). Investigating the effects of the fixed and varying dispersion parameters of Poisson-gamma models on empirical Bayes estimates. Accident Analysis & and Prevention, 40(4), 1441-1457.

Lord, D., and Mannering, F., (2010). The statistical analysis of crash-frequency data: a review and assessment of methodological alternatives. Transportation Research Part A: Policy and Practice, 44(5), 291-305.

Lovegrove, G. R., and Sayed, T., (2006). Macro-level collision prediction models for evaluating neighbourhood traffic safety. Canadian Journal of Civil Engineering, 33(5), 609-621.

Lovegrove, G., Lim, C., and Sayed, T., (2009). Community-based, macrolevel collision prediction model use with a regional transportation plan. Journal of Transportation Engineering, 136(2), 120-128.

Martin, W. A., and McGuckin, N. A., (1998). Travel estimation techniques for urban planning (Vol. 365). Washington, DC: National Academy Press.

Mannering, F. L., Shankar, V., and Bhat, C. R., (2016). Unobserved heterogeneity and the statistical analysis of highway accident data. Analytic Methods in Accident Research, 11, 1-16.

Mennis, J. 2006., Mapping the results of geographically weighted regression. The Cartographic Journal, 43(2), 171-179.

Moeinaddini, M., Asadi-Shekari, Z., Sultan, Z., and Shah, M. Z., (2015). Analyzing the relationships between the number of deaths in road accidents and the work travel mode choice at the city level. Safety Science, 72, 249-254.

Nakaya, T., Fotheringham, A.S., Brunsdon, C., and Charlton, M., (2005). Geographically weighted Poisson regression for disease association mapping. Statistics in. Medicine. 24, 2695–2717.

Nelder, J. A., and Wedderburn, R. W. M., (1972). Generalized linear models. Journal of the Royal Statistical Society, 135, 370-384

Neuhaus, J. M., Hauck, W. W., and Kalbfleisch, J. D., (1992). The effects of mixture distribution misspecification when fitting mixed-effects logistic models. Biometrika, 755-762.

Neuhaus, J. M., McCulloch, C. E., and Boylan, R., (2013). Estimation of covariate effects in generalized linear mixed models with a misspecified distribution of random intercepts and slopes. Statistics in Medicine, 32(14), 2419-2429.

Páez, A., Farber, S., and Wheeler, D., (2011). A simulation-based study of geographically weighted regression as a method for investigating spatially varying relationships. Environment and Planning-Part A, 43(12), 2992.

Pirdavani, A., Bellemans, T., Brijs, T., and Wets, G., (2014). Application of geographically weighted regression technique in spatial analysis of fatal and injury crashes. Journal of Transportation. Engineering, 140(8).

Rhee, K. A., Kim, J. K., Lee, Y. I., and Ulfarsson, G. F. (2016). Spatial regression analysis of traffic crashes in Seoul. Accident Analysis and Prevention, 91, 190-199.

SAS University Edition. SAS Institute Inc., Cary, NC, USA Accessed on 11th July 2016.

Shariat-Mohaymany, A., Shahri, M., Mirbagheri, B., and Matkan, A. A., (2015). Exploring spatial non - stationarity and varying relationships between crash data and related factors using geographically weighted poisson regression. Transactions in GIS, 19(2), 321-337.

Schwarz, G., (1978). Estimating the dimension of a model. The Annals of Statistics, 6(2), 461-464.

Shmueli, G., (2010). To explain or to predict? Statistical Science, 289-310.

Washington, S. P., Karlaftis, M. G., and Mannering, F. (2010). Statistical and econometric methods for transportation data analysis. CRC press.

Washington, S. P., Persaud, B. N., Lyon, C., and Oh, J. (2005). Validation of accident models for intersections (No. FHWA-RD-03-037).

Xu, P., and Huang, H., (2015). Modeling crash spatial heterogeneity: Random parameter versus geographically weighting. Accident Analysis and Prevention, 75, 16-25.

Yao, S., Loo, B. P., and Lam, W. W., (2015). Measures of activity-based pedestrian exposure to the risk of vehicle-pedestrian collisions: Space-time path vs. potential path tree methods. Accident Analysis and Prevention, 75, 320-332.

Young, J., and Park, P.Y., (2013). Benefits of small municipalities using jurisdiction-specific safety performance functions rather than the highway safety manual's calibrated or uncalibrated safety performance functions, Canadian Journal of Civil Engineering, 40 (6), 2013

Yu, D., (2010). Exploring spatiotemporally varying regressed relationships: The geographically weighted panel regression analysis. Proceedings of the Joint International Conference on Theory, Data Handling and Modelling in Geospatial Information Science.

CHAPTER 5:DEVELOPMENT OF MACRO-LEVEL CRIMES PREDICTION MODELS

5.1 Background

The study also developed a set of macro-level (zonal level, aggregate level) crime prediction models using advanced statistical modelling techniques called geographically weighted Poisson regression (GWPR) and geographically weighted negative binomial regression (GWNBR). The main purpose of the development of these statistical models was to assist in screening locations with certain types of crime. These locations were defined as hotzones. GWPR and GWNBR were selected because they can handle the common technical challenges embedded in many spatially distributed count data including crime data. The challenges are: 1) over-dispersion, 2) spatial dependency, and 3) regression-to-the-mean.

Five years (2009-2013) of violent and non-violent crime data for the City of Regina, Saskatchewan, Canada were used to demonstrate the development of spatial crime prediction models that take into account over-dispersion, spatial dependency and regression-to-the-mean.

5.2 Over-dispersion in Crime Data

Over-dispersion is probably the most common technical challenge that characterizes count data including crime data. Over-dispersion implies a greater variability in observed dependent variable than predicted by a statistical model (Berk and MacDonald, 2008; Hinde and Demetrio, 1998). When the dependent variable is a count variable, as in our case, Poisson regression may appear to be a straightforward option for estimating the response (Walters, 2007; Winkelman, 2008), but Poisson regression assumes that the mean is equal to the variance. This assumption is often too strict and leads to a biased result when modelling over-dispersed count data beyond the level that additional independent input variables (such as population in our case) can take into account (Coxe et al., 2009). This is due to the fact that, in many cases, all relevant input variables in the model cannot be collected and added. The omitted variable issue is one of the main sources of over-dispersion (Berk and MacDonald, 2008).

Negative binomial regression allows greater variability in the estimated dependent variable than does Poisson regression and thus can take over-dispersion in crime data into account (Berk and MacDonald, 2008; MacDonald and Lattimore, 2010; Osgood, 2000). As a result, negative binomial regression has become a standard model in crime modelling (Helbich et al., 2013, Vogel and South, 2016).

5.3 Spatial Dependency in Crime Data

The concept of spatial dependency was well established by Tobler (1970) in his statement that "everything is related to everything and closer things are more related than distant things." This statement implies the existence of spatial dependency in spatial data and has been regarded as the first law of geography (Cheong, 2012; Tita and Radil, 2010). In general, spatial patterns that result in geographical clusters of similar attributes can be viewed as evidence of spatial dependency in the data.

In the case of crime, certain types of criminal activities are often distributed non-randomly across a city with clusters in certain areas. The clustered locations can vary for different types of crime and may vary very considerably (Bernasco and Steenbeek, 2017; Curiel et al., 2017; De Melo et al., 2015; Eck et al., 2005; Farell, 2015; Johnson and Bower, 2014; Lee et al., 2015; Liu et al., 2016; Ratcliffe, 2010; Tompson and Uhlig 2008). Many studies, particularly in criminology, have investigated the non-random locations of clustered crimes to assist police services to develop more focused and informed preventative enforcement approaches.

The field of study known as the social ecology of crime examines the spatial relationship between a certain type of crime (e.g., violent crime) and its associated spatial characteristics (e.g., land use and population) (Andresen, 2012; Anselin et al., 2000; Gruenewald et al., 2006; Walker et al., 2014). Studies have found meaningful associations between the clustered locations of certain types of crimes and measured spatial characteristics (e.g., income level, land use and population), unobserved heterogeneity (e.g., characteristics that are unmeasured due to financial, technical or other data limitations) or both measured and unmeasured spatial characteristics (Radil, 2016). The meaningful associations are known as spatial dependency (Grubesic et al., 2014; Mazzulla and Forciniti, 2012).

Spatial dependency amongst characteristics across an area can be positive or negative. When similar values of certain characteristics form clusters, there is a positive association (Mitchell, 2013; Shimanda, 2004). When dissimilar values form clusters, there is a negative association. In crime studies, positive association is common. This is partly because of the socioeconomic characteristics distributed within adjacent areas. Residents with similar financial circumstance and cultural background and living within a certain area may create an example of positive association with a certain type of crime. For example, compared with high income areas, low income residential areas may have more crimes related to violence and residential burglaries (Levitt 1999; Nilsson and Estrada, 2006; Nilsson and Estrada, 2007; Suonpää et al., 2018). Land use can also play an important role in the formation of a positive association. For example, compared with residential areas, commercial areas may have more break and enter crimes (Radil, 2016).

Poisson regression and negative binomial regression assume that the target variable (e.g., the number of crimes per location) occurs independently in the various locations (McCullagh and

Nelder, 1989; Mohebbi et al., 2011). This assumption has been questioned by the many researchers who have recognized many crimes can be clearly associated with some or many of an area's spatial and demographic characteristics. The effect can be seen even in relatively small areas. This reality suggests that individual crime occurrences may not be completely independent from each other. In other words, the incidence of certain types of crimes occurring in adjacent areas may be affected by similar spatial characteristics in the adjacent areas.

The accuracy of estimating the number of crimes using Poisson and negative binomial modelling techniques then becomes questionable. If spatial dependency in data is not taken into account, the variance of the model parameters may be inflated, and inferences could be biased (LeSage, 2010).

5.4 Regression-to-the-mean

Police services typically use crime maps to highlight areas with a high concentration of a certain type of crime. Unfortunately, simple visual inspection of crime maps can be misleading as the determination of genuine hotzones is not straightforward. For example, if the hotzones are simply based on historical crime data, the maps will not take into account likely changes over time in the locations of hotzones. The locations can change from month to month or from year to year making the satisfactory identification of hotzones a complicated undertaking.

Changes occur due to regression-to-the mean (RTM) (Farrington and Welsh, 2006; Marchant 2004). RTM refers to a statistical tendency embedded in many count data. In this case, RTM is important because locations may experience a high number of crimes in a particular month or year and then, without outside intervention, experience a low number in the following month(s) or year(s). If the analyst is unaware of RTM, misinterpretations of the data and apparent trends are highly likely.

To avoid the pitfalls of RTM bias, police services need to develop maps based on the long term mean value of the number of crimes. The creation of these maps requires rigorous spatial data analysis based on appropriate and advanced statistical modelling techniques (Takyi et al., 2018). The empirical Bayes (EB) method is a widely used technique designed to mitigate RTM. The EB method combines two key pieces of information, the observed number and the predicted number, to produce the long term mean value that takes into account variations over time (e.g., from month to month or year to year).

5.5 Study Objectives

This study has three main objectives:

1. To develop macro-level crime prediction models based on two geographically weighted regression techniques (GWPR and GWNBR) to take into account spatial dependency and over-dispersion in crime data.
2. To evaluate the predictive performance of the GWPR and GWNBR models developed by considering fixed Gaussian bandwidth in parameter estimation; and
3. Determine hotzones for violent and non-violent crimes using GWNBR models combined with EB method to account for possible regression-to-the-mean-bias.

Section 5.6 provides a brief literature introducing the advanced statistical methods applied. Section 5.7 discusses the crime and other data used in the study. Section 5.8 describes the two statistical models applied in the study. Section 5.9 presents the results of the analysis. Section 5.10 summarizes the study and makes recommendations for future research.

5.6 Literature Review

During the last two decades, many researchers have addressed the development of better-fitting prediction models and they have considered a wide range of advanced statistical methods (Fahrmeir and Tutz, 2013).

Nakaya et al. (2005), for example, attempted to extend conventional count data analysis such as Poisson and negative binomial regression modelling to accommodate spatial dependency. The researchers introduced a new spatial count data modelling technique known as geographically weighted Poisson Regression (GWPR). GWPR took into account spatial dependency in a study that predicted the working age diseases related death using selected socioeconomic variables (e.g. home ownership, unemployment, proportion of persons over 64, and proportion of professional and technical workers). They used one year (1990) of data collected from the 262 municipalities in the Tokyo Metropolitan Area, Japan (Nakaya et al., (2005) and used municipality as the area unit of analysis.

The GWPR technique was considered advanced compared to the conventional Poisson and negative binomial model: GWPR produced different regression parameters for the input variables for the different areas within a study area whereas the conventional Poisson and negative binomial models produced a global (single) regression parameter for each input variable and this parameter was assumed to be constant for all the areas analyzed in a study area. When using GWPR, the model parameters estimated for a target area are more similar to those of adjacent areas than to those of remotely located areas in observance of Tobler's first law of geography (Mitchel 2005; Tobler 1970). GWPR has recently gained some attention in crime studies (Chen et al., 2017a; Chen et al., 2017b; Zhou et al., 2017).

Although GWPR can take into account spatial dependency, it cannot take into account the problem of over-dispersion in count data. This is because GWPR still assumes that the dependent variable follows the Poisson distribution and that the mean is the same as (or very close to) the variance.

Da Silva and Rodriguez (2014) enhanced GWPR by introducing geographically weighted negative binomial regression (GWNBR). GWNBR is designed to accommodate over-dispersion as well as spatial dependency in count data. GWNBR has been applied in transportation research (Gomes et al., 2017; Yu and Xu, 2017), but is new to crime research.

As mentioned, the EB technique is a popular approach to reducing RTM by estimating a long term mean value rather than simply using an historical value. The EB approach is widely used in collision research and has also been used in research into areas prone to both a high number of crimes and a high number of collisions (Hauer et al., 2002; Shaheed et al., 2015). The EB method is used in this study to reduce RTM bias.

5.7 Study Area and Data

The study area for our analysis was the City of Regina, the capital of the Canadian province Saskatchewan. The City's population was 253,220 in 2017 (Statistics Canada, 2018). In 2016, Regina's total crime rate and crime severity index were the highest for any large city in Canada. During the five years from 2009 to 2013, there were 125,338 criminal code crimes in Regina.

The study described in this chapter used five years (2009-2013) of Regina crime, demographic and land use data to develop crime prediction models. These data were obtained from three databases supplied by Regina agencies and from one online database. The data obtained were:

- Crime data (ASCII Text file format) from RPS;
- Socio-demographic data (Microsoft Excel format) and traffic data (GIS shapefile format) from the City of Regina; and
- Land use data (Microsoft Excel format), traffic analysis zone (TAZ) boundary information and roadway data (GIS shapefile format) from the City of Regina. The roadway data included the location of every intersection and roadway segment in the city's entire roadway network.
- Other data

Other aggregated variables provided for TAZs included the average parking cost and the VKMT. This information was supplied by SGI.

Table 5-1 in Section 5.7.4 shows the descriptive statistics used in this study.

5.7.1 Crime Data

This study included the ten types of crime considered the most important by RPS. The ten types of crime were grouped into two categories: violent (arson, assault, murder, robbery, and sexual assault), and non-violent (break and enter, mischief, theft, theft from auto, and theft of auto).

During the five-year study period, Regina recorded 125,338 crimes. This study included 50,284 of the 125,338 crimes: 9,181 violent and 41,103 non-violent. The data available for each crime included type of crime, location of crime (address and coordinate), and time of crime.

The individual crimes were initially imported and displayed using a GIS tool (ArcGIS) in an electronic map format. The individual crimes were then aggregated spatially by Regina's 244 TAZs. The aggregation addressed privacy and confidentiality issues as well as modelling issues.

Table 5-1 shows the two crime variables: Total Number of Non-Violent Crimes and Total Number of Violent Crimes. These two variables were considered the dependent variables.

5.7.2 Socio-demographic Data

The socio-demographic data included, for example, age (grouped), post-secondary enrollment and population per TAZ. Population density per TAZ was estimated using ArcGIS. Population was used as the primary exposure variable in the models developed: the relationship between population and crime is well documented in the literature which shows that the number of crimes increases as the population increases (Andresen, 2007; Appiahene-Gyamfi, 2002; Foote, 2015; Nolan 2004; Ousey, 2000; Zhong et al., 2011). Population increase has been shown to be the best predictor of violent and property crimes (Boivin, 2018; Chamlin and Cochran, 2004). Boivin, (2018), however, stated that the strength of this relationship varies due to the heterogeneous nature of count crime data.

Table 5-1 shows the nine socio-demographic variables. These variables were input variables.

5.7.3 Land Use Data

The land use data were analyzed in 15 categories describing various aspects of retail, office and industry floor space per TAZ. Traffic Analysis Zone Area was an additional variable. The area of each TAZ was estimated using ArcGIS.

Table 5-1 shows the 16 land use variables. These variables were also input variables.

5.7.4 Other Data

Three additional variables were also used: Average Parking Cost per Zone, Total Vehicle Kilometres Travelled and Person Vehicle Kilometres Travelled. Table 5-1 shows details of the three additional variables. These variables were also input variables.

Table 5-1: Descriptive Statistics for Crime, Socio-Demographic, Land Use and Additional Variables

Variable	Description	Mean	Std. Dev	Minimum	Maximum
Dependent Variables					
VIOLENT_CRIMES	Total Numbers of Violent Crimes	37.63	81.44	0.00	943.00
NON_VIOLENT_CRIMES	Total Numbers of Non-Violent Crimes	168.45	182.28	0.00	1,480.00
Independent Variables					
Socio-Demographic Variables					
POP_0to17	Proportion of Persons Age 0 to 17	0.17	0.09	0.00	0.29
POP_18to24	Proportion of Persons Age 18 to 24	0.09	0.04	0.00	0.33
POP_25to44	Proportion of Persons Age 25 to 44	0.25	0.10	0.00	0.40
POP_45to64	Proportion of Persons Age 45 to 64	0.23	0.09	0.00	0.34
POP_65plus	Proportion of Persons Age 65 plus	0.16	0.13	0.00	0.55
TOT_POP	Total Population	808.31	798.35	0.00	3,011.00
POP_DENSITY	Population Density (Persons/km^2)	2,121.49	1,673.68	0.00	10,552.61
NO_GRDSCH	Proportion of Persons Enrolled in Graduate School	0.10	0.33	0.00	3.90
NO_PSSTUD	Proportion of Post-Secondary Enrolment	0.43	5.09	0.00	73.31
Land Use Variables					
OFFICE_SPACE	Proportion of Office Space per Traffic Analysis Zone	0.06	0.21	0.00	1.90
RETAIL_SPACE	Proportion of Retail Space per Traffic Analysis Zone	0.04	0.09	0.00	0.78
INDUSTRY_SPACE	Proportion of Industry Space per Traffic Analysis Zone	0.01	0.03	0.00	0.20
HOSPT_SPACE	Proportion of Hospital Space per Traffic Analysis Zone	0.00	0.02	0.00	0.24
NO_LU_PER_TAZ	Number of Land Uses per Traffic Analysis Zone	4.22	1.65	1.00	8.00
AIRPORT_AREA	Proportion of Airport Area per Traffic Analysis Zone	0.00	0.05	0.00	0.85
COMMERCIAL_AREA	Proportion of Commercial Area per Traffic Analysis Zone	0.22	0.35	0.00	1.00

Table 5-1: Descriptive Statistics for Crime, Socio-Demographic, Land Use and Additional Variables (Cont'd)

Variables	Description	Mean	Std. Dev	Minimum	Maximum
Land Use Variables (Cont'd)					
INDUSTRIAL_AREA	Proportion of Industrial Area per Traffic Analysis Zone	0.09	0.26	0.00	1.00
INSTITUTIONAL_AREA	Proportion of Institutional Area per Traffic Analysis Zone	0.03	0.08	0.00	0.88
OPENSPACE_RECREATION_AREA	Proportion of Open space/Recreational Area per Traffic Analysis Zone	0.14	0.23	0.00	1.00
RAILWAY_AREA	Proportion of Railway Area per Traffic Analysis Zone	0.02	0.09	0.00	1.00
RESIDENTIAL_HD_AREA	Proportion of Residential High Density Area per Traffic Analysis Zone	0.04	0.08	0.00	0.62
RESIDENTIAL_LD_AREA	Proportion of Residential Low Density Area per Traffic Analysis Zone	0.34	0.33	0.00	0.97
RESIDENTIAL_MD_AREA	Proportion of Residential Medium Density Area per Traffic Analysis Zone	0.04	0.12	0.00	0.91
URBAN_HOLDING_AREA	Proportion of Urban Holding Area per Traffic Analysis Zone	0.07	0.22	0.00	1.00
TAZ_AREA	Traffic Analysis Zone Area (km^2)	0.51	0.56	0.01	6.04
Other Variables					
PARKING_COST	Average Parking Cost per Zone	0.70	1.81	0.00	5.95
TVKMT	Total Vehicle Kilometres Travelled (vehicles × kilometres)	25,034.04	21,770.69	321.9	122,800.88
PVKMT	Person Vehicle Kilometres Travelled (vehicles ×kilometres/population)	186.17	1,294.49	0.00	16,854.63

5.8 The GWPR and GWNBR Models

Two forms of geographically weighted regression models (GWPR and GWNBR) were considered. Sections 5.8.1 and 5.8.2 present the mathematical description of the GWPR and GWNBR models respectively. Section 5.8.2 also discusses bandwidth and the goodness-of-fit tests used to assess the two approaches. Section 5.8.3 discusses the empirical Bayes method. SAS (university edition) programming language was used to develop all the models described in this study.

5.8.1 Geographically Weighted Poisson Regression (GWPR)

Geographically Weighted Poisson Regression (GWPR) is an advancement of Poisson regression that allows model parameters to vary from one zone to another (in this case from TAZ to TAZ).

Equation (5-1) give the functional form of the Poisson regression method:

$$\mu_i = t_j exp\left(\sum_{i=1}^{k} \beta_i x_{ij}\right) \tag{5-1}$$

By allowing for spatial variation in coefficients μ_i, and y_i, the Poisson model given in Equation (5-1) can be written as Equations (5-2) and (5-3) respectively.

$$\mu_i = t_j exp\left(\sum_{i=1}^{k} \beta_i (u_j, v_j) x_{ij}\right) \tag{5-2}$$

$$y_i = Poisson\left[t_j exp\left(\sum_{i=1}^{k} \beta_i (u_j, v_j) x_{ij}\right)\right] \tag{5-3}$$

In Equation (5-3), t_j is the time period (offset variable) in which the crimes occur (in our case, five years). β_i, x_{ij} and (u_j, v_j) represent the local coefficients which can be different for each TAZ i.

Nakaya (2005) and Da Silva and Rodriguez (2014) provide detailed descriptions and parameterization of this model.

5.8.2 Geographically Weighted Negative Binomial Regression (GWNBR)

Geographically Weighted Negative Binomial Regression (GWNBR) models were also developed. The GWNBR model was proposed by Rodriguez and Da Silva (2014). As GWNBR can handle possible over-dispersion in count data, it GWNBR is expected to reduce prediction error and offer improved performance.

The GWNBR used in this study is an extension of the conventional negative binomial model shown in Equation (5-4). The GWNBR model given in Equation (5-5) allows local variation of coefficients:

$$\mu_i = NB\left[t_j \sum_{i=1}^{k} \beta_i x_{ij}\,, \alpha\right] \tag{5-4}$$

$$y_i = NB\left[t_j exp\left(\sum_{i=1}^{k} \beta_i\left(u_j, v_j\right) x_{ij}\right), \alpha\right] \tag{5-5}$$

The parameters in Equations (5-3) and (5-5) are the same as those in the previous Equations except that α is the dispersion parameter which is estimated at the global level in the two GWNBR models. Da Silva and Rodriguez (2014) provide a detailed description of GWNBR regression methods.

GWPR and GWNBR models need to determine optimal bandwidth to calibrate and produce the coefficients associated with the input variables. For both models, fixed Gaussian bandwidth as shown in Equation (5-6) was used:

$$w_{ij} = exp\left\{-\frac{1}{2}\left({}^{d_{ij}}/_{b}\right)^2\right\} \tag{5-6}$$

where:

w_{ij} represents the geographic weight of zone j when calibrating models for zone i;

$d_{ij}$is the distance between the j^{th} and i^{th} nearest neighbour zones (TAZs); and

b is the bandwidth.

An examination of cross-validation error was employed to determine the optimal bandwidth (Paez and Faber, 2007; Pirdivani et al., 2014; Wang et al., 2008). In all cases, the bandwidth that minimized the cross-validation error was selected. Equation (5-7) shows the cross-validation error:

$$CV = \sum_{i=1}^{n} \left[y_i - \hat{y}_{\neq i(b)} \right] \tag{5-7}$$

where:

CV is the cross-validation error;

$\hat{y}_{\neq j}(b)$ is the adjusted value to point i omitting the observation of i; and

n is the numbers of zones.

$\hat{y}_{\neq i}(b)$ is invariably the fitted value of y_i when the i^{th} zone is omitted during the calibration process.

The performance of the models developed was evaluated using selected goodness of fit (GOF) tests. Washington et al., (2005) provide a discussion of GOF tests. The seven GOF tests used in this study were: Akaike Information Criterion (AIC), Corrected Akaike Information Criteria (AICc), Bayesian Information Criterion (BIC), Mean Square Error (MSE), Mean Square Prediction Error (MSPE), Mean Prediction Bias (MPB), and Mean Absolute Deviation (MAD).

5.8.3 Empirical Bayes Method

As mentioned in Section 5.4, the EB method was used to mitigate the problem of RTM (Hauer et al., 2002). Equation (5-8) shows how information from similar TAZs was used to estimate the long term mean values for violent and non-violent crimes for each TAZ:

$$E[y_i] = w \cdot \mu_i + (1 - w) y_i \tag{5-8}$$

where:

$E[y_i]$ represents the EB adjusted value of violent crimes or non-violent crimes;

μ_i represents the predicted value of violent crimes or non-violent crimes;

y_i represents the observed value of violent crimes or non-violent crimes; and

w is the EB weight factor.

The EB weight factor is given in Equation (5-9):

$$w = \frac{1}{1+\alpha\times\sum_{t=1}^{Y}\mu_i} \tag{5-9}$$

where:

α is the dispersion parameter;

μ_i is the predicted number of violent crimes or non-violent crimes for TAZ i in year t; and

Y is the five years in our study.

5.9 Results of Analysis

This section presents results only for the independent variables that were statistically significant at the 0.05 significance level (see Appendix G).

5.9.1 Estimated Parameters

Several functional forms of crime prediction models were tested, but this study only reports the functional forms that best predicted violent and non-violent crime (see Table 5-2). As the value of each parameter can be different for each TAZ, Tables 5-3 and 5-4 show the range of parameter values found to be statistically significant in the GWPR and GWNBR models respectively. Each

Table shows the results for the fixed Gaussian bandwidth results for the total violent and total non-violent crime models.

The models calibrated using adaptive bi-square bandwidth for both violent and non-violent crimes generated highly extreme coefficients. Every coefficient obtained for the GWNBR models with adaptive bi-square bandwidth was of the order of 10^6 (see Appendix G). The reason for the large coefficients of the adaptive-bi-square bandwidth models will require further investigation which is beyond the scope of this research.

The large coefficient with adaptive bi-square bandwidth poses a limitation in the estimation of the performances of the models developed. This was due to the difficulty in determining the predicted values. Thus, it was impossible to compare the fixed Gaussian and the adaptive bi-square bandwidth GWPR and GWNBR models for violent and non-violent crimes. The comparison between the GWPR and GWNBR models discussed focuses on the fixed Gaussian bandwidth.

Table 5-2: Functional Forms for Best Predicting Models for Violent and Non-Violent Crimes

Severity	Functional Form
Violent Crimes	$\mu = exp^{\beta_0+\beta_1 \ln(TOP\_POP)+\beta_2 \times INDUSTRY\_SPACE+\beta_3 \times NO\_LU\_PER\_TAZ+\beta_4 \times COMMERCIAL\_AREA+\beta_5 \times RESIDENTIAL\_MD\_+\beta_6 \times URBAN\_HOLDING\_AREA}$
Non-Violent Crimes	$\mu = exp^{\beta_0+\beta_1 \ln(TOP\_POP)+\beta_2 \times POP_{65plus}+\beta_3 \times RETAIL_{SPACE}+\beta_4 \times INDUSTRY\_SPACE+\beta_5 \times NO\_LU\_PER\_TAZ+\beta_6 \times COMMERCIAL\_AREA+\beta_7 \times URBAN\_HOLDING\_AREA}$

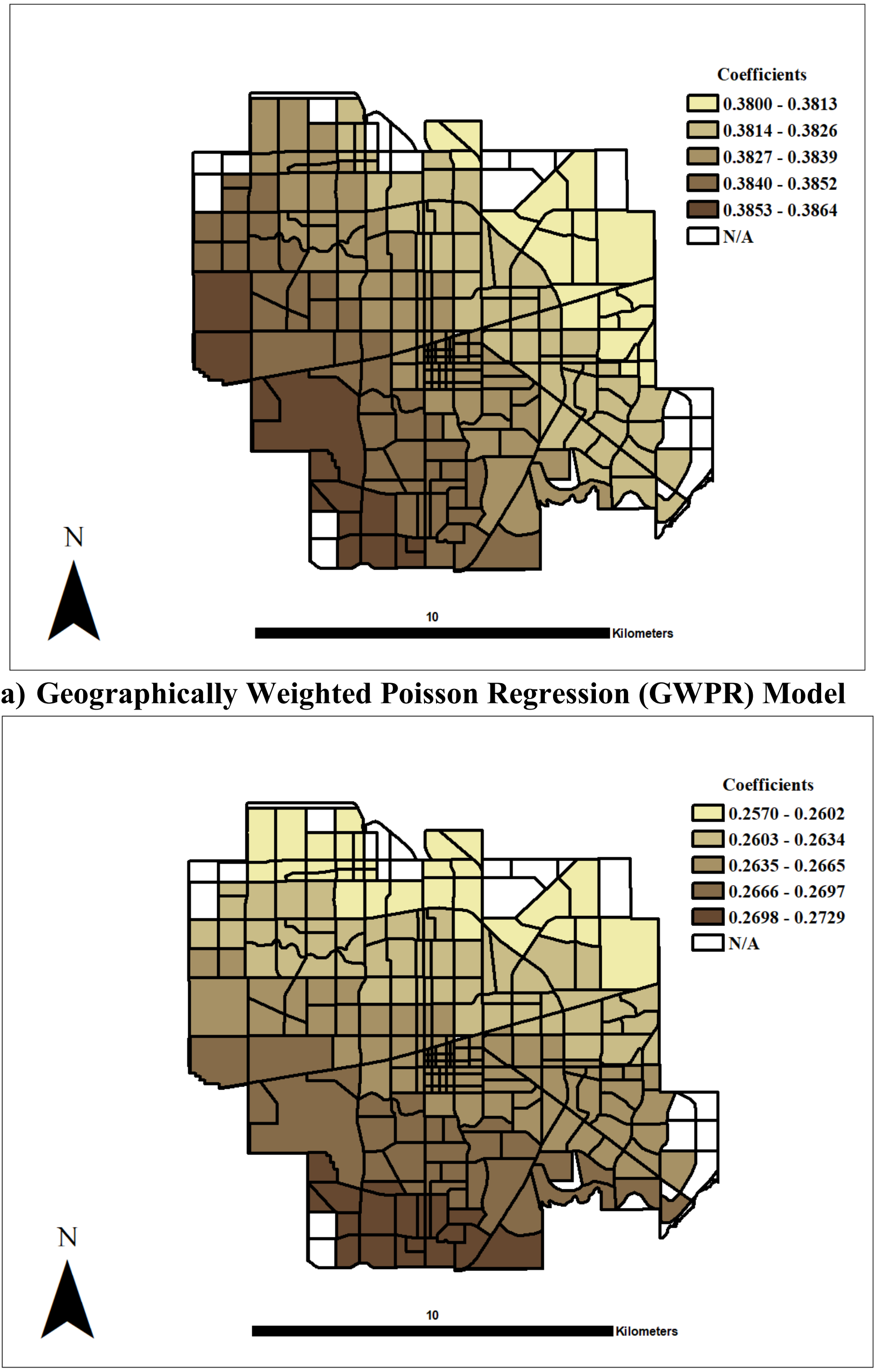
Table 5-3: Range of Parameter Values for Variables found to be Statistically Significant in the Geographically Weighted Poisson Regression (GWPR) Model

Crimes Category	Variables	Fixed Gaussian Bandwidth				
		Min	1st Quart	Median	3rd Quart	Max
Violent Crimes	Intercept	-1.662	-1.640	-1.634	-1.627	-1.607
	LOG_TOT_POP	0.380	0.382	0.383	0.384	0.386
	INDUSTRY_SPACE	8.047	8.252	8.384	8.458	8.691
	NO_LU_PER_TAZ	0.230	0.232	0.233	0.234	0.237
	COMMERCIAL_AREA	0.639	0.646	0.650	0.653	0.661
	RESIDENTIAL_MD_AREA	1.674	1.690	1.700	1.708	1.730
	URBAN_HOLDING_AREA	-6.426	-6.330	-6.296	-6.256	-6.159
Non-Violent Crimes	Intercept	0.765	0.941	1.046	1.089	1.199
	LOG_TOT_POP	0.242	0.261	0.275	0.284	0.319
	POP_65plus	-2.011	-1.923	-1.845	-1.551	-0.672
	RETAIL_SPACE	2.600	3.062	3.232	3.299	3.560
	INDUSTRY_SPACE	6.608	7.804	8.209	8.853	10.357
	NO_LU_PER_TAZ	0.153	0.201	0.211	0.225	0.241
	COMMERCIAL_AREA	0.118	0.230	0.258	0.305	0.559
	URBAN_HOLDING_AREA	-5.723	-5.138	-4.986	-4.344	-3.217

Table 5-4: Range of Parameter Values for Variables found to be Statistically Significant in the Geographically Weighted Negative Binomial Regression (GWNBR)

Crimes Category	Variables	Fixed Gaussian Bandwidth				
		Min	1st Quart	Median	3rd Quart	Max
Violent Crimes	Intercept	-1.337	-1.300	-1.286	-1.271	-1.249
	LOG_TOT_POP	0.257	0.262	0.264	0.266	0.273
	INDUSTRY_SPACE	8.538	9.004	9.252	9.431	10.107
	NO_LU_PER_TAZ	0.290	0.298	0.299	0.303	0.309
	COMMERCIAL_AREA	0.764	0.782	0.796	0.807	0.839
	RESIDENTIAL_MD_AREA	1.984	2.047	2.078	2.104	2.184
	URBAN_HOLDING_AREA	-3.945	-3.816	-3.760	-3.721	-3.575
Non-Violent Crimes	Intercept	0.783	0.855	0.876	0.889	0.920
	LOG_TOT_POP	0.236	0.244	0.250	0.252	0.262
	POP_65plus	-1.347	-1.284	-1.251	-1.167	-0.970
	RETAIL_SPACE	2.971	3.082	3.130	3.158	3.253
	INDUSTRY_SPACE	8.433	8.993	9.209	9.420	10.224
	NO_LU_PER_TAZ	0.218	0.231	0.237	0.246	0.261
	COMMERCIAL_AREA	0.474	0.500	0.518	0.531	0.587
	URBAN_HOLDING_AREA	-3.348	-3.268	-3.228	-3.181	-3.056

Figure 5-1 is an example of the distribution of estimated coefficients. The Figure shows the distribution of estimated coefficients for LOG_TOT_POP across Regina TAZs for total violent crimes. Figure 5-1a) shows the GWPR model and Figure 5-1b) shows the GWNBR model.



(a) Geographically Weighted Poisson Regression (GWPR) Model

(b) Geographically Weighted Negative Binomial Regression (GWNBR) Model

Figure 5-1: Distribution of Estimated Coefficients for LOG_TOT_POP for Total Violent Crimes

Although the main purpose of developing the models was to screen for hotzones, several points were observed.

Total population (LOG_TOT_POP) (the study's exposure variable) had a positive association with both the number of violent crimes and the number of non-violent crimes. This means that as the total population increased so did the number of violent and non-violent crimes. Similarly, both the proportion of industry space (INDUSTRY_SPACE) and the proportion of commercial area (COMMERCIAL_AREA) were associated with an increasing number of violent and non-violent crimes.

It was also found that the number of land uses per TAZ (NO_LU_PER_TAZ), which represents the different levels of complexity in land use, was associated with an increasing number of violent and non-violent crimes, and the proportion of retail space (RETAIL_SPACE) was associated with an increasing number of non-violent crimes. The number of violent and non-violent crimes decreased as the proportion of urban holding areas (URBAN_HOLDING_AREA) in a TAZ increased. This result could be expected as urban holding areas are the open space areas designated for future development (i.e., undeveloped areas).

Medium density residential areas (RESIDENTIAL_MD_AREA) showed a positive association with violent crimes. An increase in the area of medium density residential development was associated with an increase in the number of violent crimes. This result could be expected as the density of residential areas has been reported to influence violent crimes such as aggravated assault and robbery (Browning et al., 2010).

The proportion of persons age 65 plus (POP_65plus) was negatively associated with non-violent crimes. This result could be expected as seniors are known to have lower crime rates than certain younger groups.

5.9.2 Model Selection

Seven GOF tests listed in Section 5.8.2 and discussed in chapter 3 were used to select the best fitting model for total violent and total non-violent crimes. Table 5-5 shows the results. Smaller values in the GOF tests indicate a better fitting model. The cells shaded grey highlight the present better fitting results when comparing the GWPR and GWNBR models.

Table 5-5: Results of Goodness of Fit Tests for Total Violent and Total Non-Violent Crime Models

Crime Category	Model	Dispersion	AIC	AICc	BIC	MSE	MSPE	MPB	MAD
Violent Crimes	**GWPR**	-	12306.95	12307.48	12332.77	**232.33**	**225.67**	-0.09	6.16
	GWNBR	1.00	**2064.75**	**2065.45**	**2094.58**	241.52	234.59	**-0.07**	**6.15**
Non-Violent Crimes	**GWPR**	-	18188.32	18190.24	18238.34	**785.37**	**759.62**	-1.60	**16.70**
	GWNBR	0.49	**2781.25**	**2782.35**	**2819.11**	882.52	853.58	**-0.81**	17.18

Table 5-5 shows that:

- The GOF tests did not consistently favour the models for violent crimes and non-violent crimes or the GWPR or GWNBR models. Inconsistency in GOF tests has previously been reported by Park and Young (2013);
- In the case of the violent crime models, five of the seven GOF tests (AIC, AICc, BIC, MPB, and MAD) favoured the GWNBR model and two tests (MSE and MSPE) favoured the GWPR model.

- In the case of the non-violent crime models, four GOF tests (AIC, AICc, BIC, and MPB) favoured the GWNBR model and three tests (MSE, MSPE and MAD) favoured the GWPR model.

After examining the GOF results, the GWNBR was selected as the better fitting model for both the total violent and total non-violent crime data. GWNBR also has the two previously mentioned important advantages of being able to handle over-dispersion and being suitable for use with the EB method.

5.9.3 Hotzone Selection

The GWNBR models coupled with the EB method was then used to identify Regina's violent and non-violent crime hotzones. Figure 5-2 shows the top ten hotzones for violent and non-violent crime. The zones do not reflect simply historical data, but take into account the long term mean values for the number of violent and non-violent crimes.

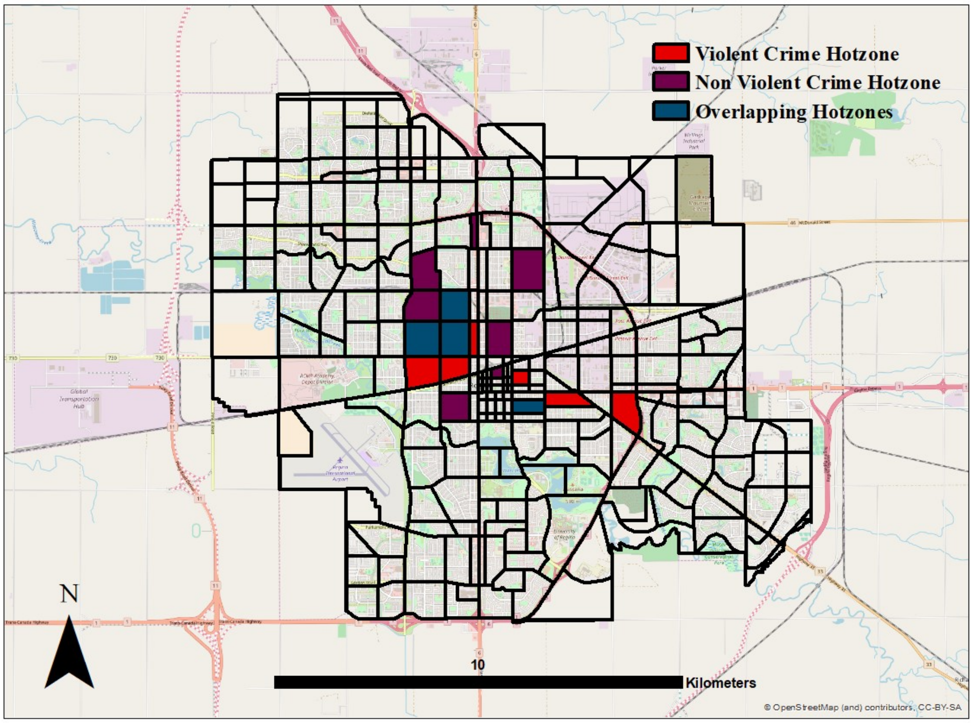

Violent Crime Hotzone
Non Violent Crime Hotzone
Overlapping Hotzones
N
10
Kilometers
© OpenStreetMap (and) contributors, CC-BY-SA

Figure 5-2: Top 10 Hotzones for Violent and Non-Violent Crime in Regina

Examination of Regina's Top 10 hotzones for violent and non-violent crime showed 4 overlapping hotzones. The overlapping hotzones are areas of low-income population or areas near downtown with mainly commercial land use. The overlapping hotzones could become areas where police could pro-actively focus their law enforcement strategies to prevent future crimes.

5.10 Summary and Recommendations for Future Research

The research described in this chapter developed and evaluated crime prediction models for the City of Regina using a GWPR approach and a GWNBR combined with EB approach. Five years (2009-2013) of crime and other data from Regina were used in the model development. The methodologies were chosen to overcome three technical challenges in count data: over-dispersion, spatial dependency and RTM.

Both the GWPR and GWNBR models took the spatial dependency issue into account and both successfully produced spatially varying parameters for each input variable. The GWNBR model was able to estimate the over-dispersion parameter to take into account over-dispersed count data. Due to the GWPR and GWNBR models requiring separate regression equations for all data points, partitioning the data into calibration and validation datasets were a challenge. Thus, the same data used for calibrating the models was used as a validation dataset for obtaining the GOF of the models that depends on the residuals.

The performance of the models developed was compared using seven GOF tests (AIC, AICc, BIC MSE, MSPE, MPB, and MAD). The GOF test results were inconclusive and caution is recommended when interpreting the results of GOF tests especially if only a few tests are used.

In this study, GWNBR was selected as the better model. An important benefit was the ability to couple the GWNBR model to the EB technique to reduce RTM bias by estimating the long term mean values for the number of crimes. This approach is very much preferable to relying simply on the selection of hotzones based on historical data which is likely to result in RTM bias.

The results of the GWNBR model with the EB technique were used to map the Top 10 violent and non-violent crime hotzones in Regina. These areas can be regarded as areas where security is a concern and where enforcement by police services should be focused.

The dispersion parameter in our GWNBR model was fixed. Future research could evaluate whether the model could be improved by allowing dispersion parameters to vary. Varying dispersion parameters might improve GOF results.

The Gaussian bandwidth method in the GWNBR model was also fixed. Future research could consider a different method for determining optimal bandwidth and whether the bandwidth determined would improve the result. The adaptive bi-square bandwidth method is an example of a different method (Fotheringham, 2002; Jacquez, 2010).

References

Andresen, M. A. (2007). Location quotients, ambient populations, and the spatial analysis of crime in Vancouver, Canada. Environment and Planning A, 39(10), 2423-2444.

Anselin, L., Cohen, J., Cook, D., Gorr, W., and Tita, G. (2000). Spatial analyses of crime. Criminal Justice, 4(2), 213-262.

Andresen, M. A. (2012). Unemployment and crime: A neighborhood level panel data approach. Social Science Research, 41(6), 1615-1628.

Appiahene-Gyamfi, J. (2002). An analysis of the broad crime trends and patterns in Ghana. Journal of criminal justice, 30(3), 229-243.

Bernasco, W., and Steenbeek, W. (2017). More places than crimes: implications for evaluating the law of crime concentration at place. Journal of Quantitative Criminology, 33(3), 451-467.

Berk, R., and MacDonald, J. M., (2008). Overdispersion and Poisson regression. Journal of Quantitative Criminology, 24(3), 269-284

Boivin, R. (2018). Routine activity, population(s) and crime: Spatial heterogeneity and conflicting Propositions about the neighborhood crime-population link. Applied Geography, 95, 79-87.

Browning, C. R., Byron, R. A., Calder, C. A., Krivo, L. J., Kwan, M. P., Lee, J. Y., and Peterson, R. D. (2010). Commercial density, residential concentration, and crime: Land use patterns and violence in neighborhood context. Journal of Research in Crime and Delinquency, 47(3), 329-357.

Chamlin, M. B., and Cochran, J. K. (2004). An excursus on the population size-crime relationship. W. Criminology Rev., 5, 119.

Chen, J., Liu, L., Zhou, S., Xiao, L., Song, G., and Ren, F. (2017a). Modeling spatial effect in residential burglary: A case study from ZG city, China. ISPRS International Journal of Geo-Information, 6(5), 138.

Chen, J., Liu, L., Zhou, S., Xiao, L., and Jiang, C. (2017b). Spatial variation relationship between floating population and residential burglary: A case study from ZG, China. ISPRS International Journal of Geo-Information, 6(8), 246.

Cheong, J. (2012). The effect of neighborhood disorder on crime: a spatial analysis. Statistics (Journal of the Korean Official Statistics), 17, 122-142.

Coxe, S., West, S. G., and Aiken, L. S. (2009). The analysis of count data: A gentle introduction to Poisson regression and its alternatives. Journal of Personality Assessment, 91(2), 121-136.

Curiel, R. P., Delmar, S. C., and Bishop, S. R. (2017). Measuring the distribution of crime and Its Concentration. Journal of Quantitative Criminology, 1-29.

Da Silva, A. R., and Rodrigues, T. C. V. (2014). Geographically weighted negative binomial regression—incorporating overdispersion. Statistics and Computing, 24(5), 769-783.

De Melo, S. N., Matias, L. F., and Andresen, M. A. (2015). Crime concentrations and similarities in spatial crime patterns in a Brazilian context. Applied Geography, 62, 314-324.

Eck, J., Chainey, S., Cameron, J., and Wilson, R. (2005). Mapping crime: Understanding hotspots.

Farber, S., and Páez, A. (2007). A systematic investigation of cross-validation in GWR model estimation: empirical analysis and Monte Carlo simulations. Journal of Geographical Systems, 9(4), 371-396.

Fahrmeir, L., and Tutz, G. (2013). Multivariate statistical modelling based on generalized linear models. Springer Science and Business Media.

Farrington, D. P., and Welsh, B. C. (2006). How important is “Regression to the Mean” in Area-Based Crime Prevention Research? Crime Prevention and Community Safety, 8(1), 50-60.

Farrell, G. (2015). Crime concentration theory. Crime Prevention and Community Safety, 17(4), 233-248.

Fotheringham, A.S., C. Brunsdon, and Charlton M.E. (2002). Geographically weighted regression: The analysis of spatially varying relationships. Chichester: Wiley.

Foote, A. (2015). Decomposing the effect of crime on population changes. Demography, 52(2), 705-728.

Gomes, M. J. T. L., Cunto, F., and da Silva, A. R. (2017). Geographically weighted negative binomial regression applied to zonal level safety performance models. Accident Analysis and Prevention, 106, 254-261.

Grubesic, T. H., Wei, R., and Murray, A. T. (2014). Spatial clustering overview and comparison: Accuracy, sensitivity, and computational expense. Annals of the Association of American Geographers, 104(6), 1134-1156.

Gruenewald, P. J., Freisthler, B., Remer, L., LaScala, E. A., and Treno, A. (2006). Ecological models of alcohol outlets and violent assaults: crime potentials and geospatial analysis. Addiction, 101(5), 666-677.

Hauer, E., Harwood, D., Council, F., and Griffith, M. (2002). Estimating safety by the empirical Bayes method: A tutorial. Transportation Research Record: Journal of the Transportation Research Board, (1784), 126-131.

Helbich, M., Jokar-Arsanjani, J., and Leitner, M. (2013). Driving forces of non-violent crime in Houston, TX: A spatially filtered negative binomial model.

Hinde, J., and Demétrio, C. G. (1998). Overdispersion: models and estimation. Computational Statistics and Data Analysis, 27(2), 151-170.

Jacquez, G. M. (2010). Space-time intelligence system software for the analysis of complex systems. In Handbook of Applied Spatial Analysis (pp. 113-124). Springer, Berlin, Heidelberg.

Johnson S.D., and Bowers K.J. (2014). Near repeats and crime forecasting. In: Bruinsma G., Weisburd D. (Eds) Encyclopedia of Criminology and Criminal Justice. Springer, New York, NY

Lee, Y., Eck, J. E., SooHyun, O., and Martinez, N. N. (2017). How concentrated is crime at places? A systematic review from 1970 to 2015. Crime Science, 6(1), 6.

LeSage, J., and Pace, R. K. (2009). Introduction to spatial econometrics. Chapman and Hall/CRC.

Levitt, S. D. (1999). The changing relationship between income and crime victimization.

Liu, D., Song, W., and Xiu, C. (2016). Spatial patterns of violent crimes and neighborhood characteristics in Changchun, China. Australian and New Zealand Journal of Criminology, 49(1), 53-72.

MacDonald, J. M., and Lattimore, P. K. (2010). Count models in criminology. In Handbook of quantitative criminology (pp. 683-698). Springer, New York, NY.

Marchant, P. R. (2004). A demonstration that the claim that brighter lighting reduces crime is unfounded. British Journal of Criminology, 44(3), 441-447.

Mazzulla, G., and Forciniti, C. (2012). Spatial association techniques for analyzing trip distribution in an urban area. European Transport Research Review, 4(4), 217-233.

McCullagh, P., and Nelder, J. A. (1989). Generalized linear models (Vol. 37). CRC press.

Mitchel, A. (2005). The ESRI guide to GIS analysis, Volume 2: Spatial measurements and statistics. ESRI Guide to GIS analysis

Mitchell, W. F. (2013). Introduction to spatial econometric modelling. Centre of Full Employment and Equity, University of Newcastle.

Mohebbi, M., Wolfe, R., and Jolley, D. (2011). A Poisson regression approach for modelling spatial autocorrelation between geographically referenced observations. BMC medical research methodology, 11(1), 133.

Nakaya, T., Fotheringham, A. S., Brunsdon, C., and Charlton, M. (2005). Geographically weighted Poisson regression for disease association mapping. Statistics in Medicine, 24(17), 2695-2717

Nilsson, A., and Estrada, F. (2006). The inequality of victimization: trends in exposure to crime among rich and poor. European journal of criminology, 3(4), 387-412.

Nilsson, A., and Estrada, F. (2007). Risky neighbourhood or individuals at risk? The significance of neighbourhood conditions for violent victimization in residential areas. Journal of Scandinavian studies in criminology and crime prevention, 8(1), 2-21.

Nolan, J. J. (2004). Establishing the statistical relationship between population size and UCR crime rate: Its impact and implications. Journal of Criminal Justice, 32(6), 547-555.

Osgood, D. W. (2000). Poisson-based regression analysis of aggregate crime rates. Journal of Quantitative Criminology, 16(1), 21-43.

Ousey, G. C. (2000). Explaining regional and urban variation in crime: a review of research. Criminal justice, 1, 261-308.

Pirdavani, A., Bellemans, T., Brijs, T., and Wets, G. (2014). Application of geographically weighted regression technique in spatial analysis of fatal and injury crashes. Journal of Transportation Engineering, 140(8), 04014032.

Radil, S. M. (2016). Spatial analysis of crime. The Handbook of Measurement Issues in Criminology and Criminal Justice, 535-554.

Ratcliffe, J. (2010). Crime mapping: spatial and temporal challenges. In Handbook of Quantitative Criminology (pp. 5-24). Springer, New York, NY.

Shaheed, M. S., Gkritza, K., Hallmark, S. L., and Knapp, K. K. (2015). An Application of the empirical Bayes method for identifying winter weather crash hot spots. In Transportation Research Board 94th Annual Meeting, No. 15-4785.

Shimada, T. (2004). Spatial diffusion of residential burglaries in Tokyo: Using exploratory spatial data analysis. Behaviormetrika, 31(2), 169-181.

Statistics Canada. Annual demographic estimates: subprovincial areas, July, 1, 2017. Catalogue No. 91-214-X, 2018. ISSN 1920-8154

Suonpää, K., Kivivuori, J., and Aaltonen, M. (2018). Criminal history and social disadvantage as predictors of the severity of violent offending. International journal of comparative and applied criminal justice, 42(2-3), 139-155.

Takyi, E. A., Oluwajana, S. D., and Park, P. Y. (2018). Development of macro-level crime and collision prediction models to support Data-Driven Approach to Crime and Traffic Safety (DDACTS). Transportation Research Record, 0361198118777356.

Tita, G. E., and Radil, S. M. (2010). Spatial regression models in criminology: Modeling social processes in the spatial weights matrix. In Handbook of quantitative criminology (pp. 101-121). Springer, New York, NY.

Tobler, W. R. (1970). A computer movie simulating urban growth in the Detroit region. Economic Geography, 46(sup1), 234-240.

Tompson, L., and Townsley, M. (2010). (Looking) Back to the future: using space—time patterns to better predict the location of street crime. International Journal of Police Science and Management, 12(1), 23-40.

Vogel, M., and South, S. J. (2016). Spatial dimensions of the effect of neighborhood disadvantage on delinquency. Criminology, 54(3), 434-458.

Walters, G. D. (2007). Using Poisson class regression to analyze count data in correctional and forensic psychology: A relatively old solution to a relatively new problem. Criminal Justice and Behavior, 34(12), 1659-1674.

Walker, W. C., Sim, S., and Keys-Mathews, L. (2014). Use of geographically weighted regression on ecology of crime, response to hurricane in Miami, Florida. In Forensic GIS (pp. 245-262). Springer, Dordrecht.

Wang, N., Mei, C. L., and Yan, X. D. (2008). Local linear estimation of spatially varying coefficient models: an improvement on the geographically weighted regression technique. Environment and Planning A, 40(4), 986-1005.

Washington, S. P., Persaud, B. N., Lyon, C., and Oh, J. (2005). Validation of accident models for intersections (No. FHWA-RD-03-037).

Weisburd, D. (2015). The law of crime concentration and the criminology of place. Criminology, 53(2), 133-157.

Winkelmann, R. (2008). Econometric analysis of count data. Springer Science and Business Media

Young, J., and Park, P. Y. (2013). Benefits of small municipalities using jurisdiction-specific safety performance functions rather than the Highway Safety Manual's calibrated or uncalibrated safety performance functions. Canadian Journal of Civil Engineering, 40(6), 517-527.

Yu, C. Y., and Xu, M. (2017). Local variations in the impacts of built environments on traffic Safety. Journal of Planning Education and Research, 0739456X17696035.

Zhong, H., Yin, J., Wu, J., Yao, S., Wang, Z., Lv, Z., and Yu, B. (2011). Spatial analysis for crime pattern of metropolis in transition using police records and GIS: a case study of Shanghai, China. International Journal of Digital Contents Technology and its Applications, 5(2), 93-105.

Zhou S., Xie W., Song, G., and Liu L., (2017). The spatial differentiation effect of land use on street robbery: A case study in H city, China. Scientia Geographica Sinica, 37(6): 885-894.

CHAPTER 6:MACRO-LEVEL PREDICTION OF OVERLAPPING CRIME AND COLLISION HOTZONES FOR FOCUSED LAW ENFORCEMENT

6.1 Background

The research discussed in this dissertation was conducted for the City of Regina, the capital of Saskatchewan, Canada. In terms of population, the city is the second largest in the province with a population of 253,220 in 2017 (Statistics Canada, 2018). The city is known to have a high number of crimes and a high number of collisions relative to many other Canadian cities (Statistics Canada, 2016).

The goal of this study is to help the City of Regina, Saskatchewan to improve the effectiveness and efficiency of the city's law enforcement efforts by directing law enforcement resources to areas where a high number of crimes and a high number of collisions overlap. To meet this goal, the study identifies high crime and collision areas, or hotzones, by developing a sophisticated macro-level (zonal-level) model designed to predict crime and collision numbers.

The study's macro-level approach to collision prediction is important because of confidentiality and privacy issues associated with crime data. As it is often not possible to disclose the specific location (e.g., house) of a specific crime (e.g., murder), many studies of crime have taken a macro-level approach (Osgood, 2000; Pratt, 2010). A macro-level approach to collision prediction is also very common in transportation engineering studies (Hadayeghi et al., 2003; Hakim et al., 1991; Lovegrove and Sayed, 2006; Lovegrove et al., 2009).

This study used traffic analysis zones (TAZs) as the main spatial unit of analysis. Other possibilities included neighbourhood, county, etc. but TAZs were the obvious choice because

Regina uses TAZs as the basis for its collection of land use, population and many other spatial characteristics. The City of Regina has 244 TAZs.

6.2 Study Objectives

The study has two specific objectives:

- Develop sophisticated macro-level (zonal-level) crime and collision prediction models based on an advanced statistical technique; and
- Demonstrate how the models developed can be used to identify and map hotzones where significant numbers of total crimes and total collisions are expected to overlap.

6.3. DDACTS Background

Urbanization has generally improved our standard of living and quality of life thanks to the concentration of economic development and opportunities. Urbanization is also, however, often associated with a concentration of certain types of crime in certain urban areas (e.g., commercial areas) and a concentration of collisions on certain types of roadway (e.g., arterials) (Soh, 2012; Wegman, 2017; Wiebe et al., 2016). Residential areas and local roads are typically less affected by crime and collisions (Anderson et al., 2013; Kaygisiz et al., 2017; Lovegrove and Sayed, 2006).

The high number of crimes and collisions in urban areas is an important issue for law enforcement agencies, transportation safety professionals and city politicians. Public demands for increased law enforcement increase the cost of police services (Fell, 2013; Hardy, 2010). Law enforcement agencies need to allocate their limited financial and other resources as effectively and efficiently as possible while ensuring that they minimize the number of crimes and collisions. To achieve this goal, North American law enforcement agencies have adopted various approaches (Leigh et al., 2016, Leigh et al., 2017).

Data-Driven Approaches to Crime and Traffic Safety (DDACTS) is a good example. Police services in the United States and Canada have been using DDACTS to develop operational level enforcement tactics since 2008 (Cohen, 2014; Cook, 2012; NHTSA, 2014; Ryderberg et al, 2014; Ryderberg et al, 2018; Shimko, 2013; Takyi et al., 2015). DDACTS uses highly visible traffic enforcement to simultaneously improve both public security and traffic safety in selected target areas (hotzones). Hotzones are areas where the expected number of both crimes and collisions is higher than in other areas (Hardy, 2010; Fell, 2013; Weiss, 2013). Identifying the hotzones for focused law enforcement presents various challenges, but the rationale for the DDACTS approach is well established.

An additional important advantage to adopting a data driven approach is that a neutral and impartial scientific procedure obviously offers advantages over an ad hoc or anecdotal approach. The selection of certain areas for focused law enforcement may otherwise appear biased and unacceptable to some citizens.

6.4. Empirical Rationale for DDACTS

Criminologists have long investigated the relationship between crime and traffic collisions/violations. An early study (Michalowski, 1975) investigated 119 traffic collisions that led to one or more deaths. The study showed that the drivers involved in fatal collisions, like violent crime offenders, exhibited a tendency towards aggressive behavior.

Fleiter et al., (2015) argued that drivers who exhibit aggression while driving are more likely than other drivers to be involved in a criminal activity. The Fleiter et al. study supports the theory that individuals who have behavioural problems while driving are also likely to have behavioural problems that may be associated with other criminal activities. Other studies (Brace et al., 2009; Dodson et al., 2011; Junger et al., 2001; Sansone et al., 2011; Watson et al., 2015) have taken the view that individuals involved in certain criminal activities are more likely to be involved in severe collisions.

Catillo-Manzano et al. (2015) showed a spatial relationship between crime and collisions. They developed statistical models that used the crime rate to predict the number of fatal collisions for 28 European countries. The study was based on two years of crime and collision data (1999 and 2000).

Environmental criminology studies of deviant place theory include research that has shown that some areas within a city appear to attract individuals who commit crimes and traffic violations (some of which could lead to traffic collisions) disproportionately (Stark, 1987; Siegel and McCormick, 2009). The existence of such places suggests that crime and collision locations show concentrations in certain areas that should be the focus of increased law enforcement. Such areas are typically rough neighbourhoods with high crime rates. Deviant place theory presents the idea that individuals who have a higher exposure to risky places will have an increased likelihood of being involved in crime.

The variable vehicle kilometers travelled (VKMT) has often been used to assess exposure in collision prediction models (Aguero-Valverde, 2013; Cui et al., 2015; Hadayeghi et al, 2003; Hadayeghi et al, 2007; Hadayeghi et al., 2010; Khondakar et al., 2010; Lovegrove and Sayed, 2007; Pirdavani et al., 2014, Rhee et al., 2016). In the case of crime prediction models, the variable total population has been used as a proxy (or pseudo) exposure variable (Andresen, 2005; Andresen, 2007; Appiahene-Gyamfi; 2002; Harries, 2006; Taylor et al., 2015). Interestingly, total population is also sometimes used as a proxy (or pseudo) exposure variable for collision prediction (Ladron de Guevara et al., 2004; Noland and Quddus, 2003; Siddiqui et al., 2012). The use of total population provides interesting indirect evidence suggesting that areas of high crimes and high collisions may overlap.

Kuo et al. (2013) showed that there is a spillover effect from DDACTS zones to nearby zones. This effect adds to the potential benefits of adopting a DDACTS approach. It is clear that research has shown the rationale for law enforcement tactics, such as DDACTS, that aim to reduce the number of crimes and collisions simultaneously by focusing on certain carefully selected areas. The size of the area (e.g., zone, city, region) analyzed and the size of the hotzones identified vary and depend on the scope and purpose of the study.

6.5 Impact of DDACTS

In 1997, police services in Albuquerque, New Mexico introduced a version of DDACTS called the "Safe Streets" program. The program introduced various traffic enforcement tactics including saturation patrols, follow-up patrols, highway speed enforcement, and sobriety checkpoints. They focused their traffic enforcement on 27 high collision locations in four areas within the city boundary. The four areas were also known to be high crime areas in the City. The program resulted in an 18% decline in injury collisions, a 20% decline in impaired driving collisions, a 34% decline in fatal collisions, a 29% decline in homicides, a 17% decline in kidnapping, and a 10% decline in assaults (Stuster, 2001).

Other jurisdictions that have successfully applied DDACTS include metropolitan areas, such as Baltimore County, Maryland and Nashville, Tennessee and small towns such as St. Albans, Vermont and Sheboygan, Wisconsin (Wisconsin DOT, 2010). In the case of Baltimore County, for example, burglaries fell by 17%, motor vehicle theft by 41%, and fatal and injury collisions by 43% after the introduction of DDACTS (Hall and Puls, 2010).

6.6 DDACTS Methodologies

Most DDACTS studies have identified hotzones using a GIS mapping technique coupled with kernel density estimation (KDE) (Braga, 2006; Braga et al., 2014). This approach uses historical crime and collision data to identify existing hotzones, but we cannot assume that the same areas will remain hotzones in the future. This is because crimes and collisions could be displaced to other areas. To avoid or minimize this problem, it is important to account for spatial dependency when analyzing the data.

Crime and collision patterns are known to be associated with spatial characteristics such as the socio-demographic, land use, traffic, and surface infrastructure of a city (Wedagama et al., 2006; Weir et al., 2009; Cottrill, 2010; Walters, 2006; Troy and Grove, 2008; Wolfe and Groove, 2012; Anderson et al, 2012; Frazer et al., 2013; Song et al., 2015). As these spatial characteristics change over time, it is reasonable to expect that hotzone locations will also change over time. Law enforcement aims to prevent future crimes and collisions and therefore needs a tool that can incorporate spatial characteristics and predict future crime and collision hotzones. Such an approach supports police services' desired commitment to proactive rather than reactive crime and collision prevention.

The standard approach to estimating the number of future collisions is a conventional negative binomial model (Aguero-Valverde and Jovanis, 2006, Highway Safety Manual, 2010, Takyi et al., 2018). This method has also been used to estimate the number of crimes (Berk and MacDonald, 2008; Haining, et al., 2009; Piza, 2012; Tseloni, 2006).

Conventional negative binomial models assume that observations are independent from each other. In our case, this assumption requires us to assume that crimes and collisions occur independently of each other. The assumption that observations are independent from each other has long been questioned. Tobler (1970) introduced the well-established first law of geography: "*...everything is related to everything else, but near things are more related than distant things...*" to explain the concept of spatial correlation among study data variables. Some degree of spatial correlation is certainly to be expected among the crime, collision and other spatial issues relevant to a DDACTS study.

Yu (2010) used a new technique called geographically weighted regression (GWR) which can consider the spatial correlation in input data. GWR, however, assumes that the error-term is normally distributed. This assumption is not acceptable for count data such as crime and collision observations.

Da Silva and Rodrigues (2014) introduced a geographically weighted negative binomial regression model (GWNBR) designed to improve handling of spatially correlated count data. GWNBR has not been used in crime prediction but has been used in collision prediction (Gomes et al., 2017; Yu and Xu, 2017).

This dissertation's research used a GWNBR model to mitigate potential spatial correlation in our crime, collision and spatial data. Interestingly, this advanced statistical technique has not been used before in crime prediction or DDACTS analysis.

Three models to predict the number of crimes (total, violent and non-violent) and three models to predict the number of collisions (total, fatal + injury (FI), and property damage only (PDO)) were initially developed. All six models used a GWNBR methodology. This study concentrates on two of the models: the total number of crimes and the total number of collisions. It is believed that these two models are sufficient for demonstrating the use of macro-level crime and collision prediction models to select appropriate hotzones for DDACTS enforcement.

The total number of crimes used is defined as the total for the ten most important types of crime as specified by the RPS: arson, assault, break and enter, mischief, robbery, sexual assault, murder, theft, theft from auto, and theft of auto. (Other law enforcement agencies may select different types of crime as the most important.)

Section 6.7 of this chapter briefly discusses the crime, collision and other data used in the study. Section 6.8 discusses the main statistical method (GWNBR) applied. Section 6.9 summarizes the results of our analysis, and Section 6.10 presents the study conclusions and suggestions for further research.

6.7 Study Data

The study used five years (2009 to 2013) of data for crime, collisions, traffic volumes, roadways, land use and socio-demographics. These data were obtained from six different databases supplied by three different agencies (RPS, SGI and the City of Regina):

- Crime data (Microsoft text file format) from RPS;
- Collision data (Microsoft Access format) from SGI;
- Traffic volume data (GIS shapefile format) from the City of Regina;
- TAZ boundary information and roadway data (GIS shapefile format) from the City of Regina. The roadway data included the location of every intersection and roadway segment in the city's entire roadway network and additional data such as the functional classification and speed limit of every roadway segment;
- Land use data (Microsoft Excel format) from the City of Regina; and
- Socio-demographic data (Microsoft Excel format) from the City of Regina.

Table 6-1 lists the two dependent variables (Total Number of Crimes and Total Number of Collisions) and 18 independent variables used in the study. The independent variables are grouped under four headings: traffic volume, roadway, land use and socio-demographics. All the input variables in Table 6-1 were aggregated by TAZ.

During the five-year study period, the Total Number of Crimes was 50,284 and the Total Number of Collisions was 26,642.

Table 6-1: Input Variables

Variables	Description	Mean	Std. Dev.	Min.	Max.
Crime					
TOTAL_CRIMES	Total Number of Crimes	206.08	252.95	0.00	2423.00
Collisions					
TOTAL_CRASHES	Total Number of Collisions	109.18	87.57	1.00	425.00
Traffic volume					
TAZ_VKMT	Vehicle Kilometers Travelled/Traffic Analysis Zone	25034.04	21770.69	321.90	122800.88
Roadways					
INT_DEN	Intersection Density (Intersections/Area)	38.13	24.40	0.00	129.20
NO_3LEGS_INT	Numbers of 3 Leg Intersections	9.66	10.10	0.00	66.00
AVE_SEGLEN	Weighted Average Segment Length	2.23	1.87	0.08	11.36
TOT_SEGLEN_DEN	Total Segment Length Density	10.53	5.20	0.82	40.76
LOCAL_ROAD_DEN	Local Road Density	5.76	4.04	0.00	15.27
ARTERIAL_LEN	Arterial Length (km)	0.57	0.57	0.00	2.62
Land use					
OFFICE_SPACE	Office Space Area as Proportion of Traffic Analysis Zone	0.06	0.21	0.00	1.90
RETAIL_SPACE	Retail Space Area as Proportion of Traffic Analysis Zone	0.04	0.09	0.00	0.78
INDUSTRY_SPACE	Industry Space Area as Proportion of Traffic Analysis Zone	0.01	0.03	0.00	0.20
NO_LU_PER_TAZ	Number of Land Uses per Traffic Analysis Zone	4.22	1.65	1.00	8.00
COMMERCIAL_AREA	Commercial Area as Proportion of Traffic Analysis Zone	0.22	0.35	0.00	1.00
INDUSTRIAL_AREA	Industrial Area as Proportion of Traffic Analysis Zone	0.09	0.26	0.00	1.00
RESIDENTIAL_MD_AREA	Residential Medium Density Area as Proportion of Traffic Analysis Zone	0.04	0.12	0.00	0.91
URBAN_HOLDING_AREA	Urban Holding Area as Proportion of Traffic Analysis Zone	0.07	0.22	0.00	1.00
TAZ_AREA	Traffic Analysis Zone Area (km^2)	0.51	0.56	0.01	6.04

Table 6-1: Input Variables (Cont'd)

Variables	Description	Mean	Std. Dev.	Min.	Max.
Socio-demographics					
TOT_POP	Total Population	808.31	798.35	0.00	3011.00
POP_65plus	Persons Age 65 plus as Proportion of Total Population	0.16	0.13	0.00	0.55
Sample Size					
N	Number of Traffic Analysis Zones	244			

6.8. Methodology

6.8.1 Geographically Weighted Negative Binomial Regression

Geographically Weighted Negative Binomial Regression (GWNBR) was the main tool used to predict the number of macro-level crimes and collisions. The GWNBR model used in this study is similar to the conventional negative binomial model. The model produces a global (fixed) dispersion parameter (α) for all TAZs (Da Silva and Rodrigues, 2014).

GWNBR models can construct different equations for each zone. The parameters estimated for the different zones can be different, but the parameters estimated for a particular zone will be more similar to the parameters estimated for nearby zones and less similar to the parameters estimated for distant zones. This is how the GWNBR model adopts Tobler (1970)'s first law of geography.

The study's GWNBR model used fixed Gaussian bandwidth to differentiate between a target zone's nearby zones and distant zones. Equation (6-1) shows the mathematical expression for the GWNBR model:

$$y_i \sim NB\left[t_j\, exp\left(\sum_{i=1}^{k} \beta_i\left(u_j, v_j\right) x_{ij}\right), \alpha\right] \qquad \textit{(6-1)}$$

where:

t_j is an offset variable representing the exposure time (in this case 5 years);

β_i is the parameter of input variable x_i for $i = 1, \ldots, k$;

y_i is the ith dependent variable for zone $i = 1, \ldots, \ldots, n$; and

(u_j, v_j) represents the location coordinates for zone i.

Equation (6-2) shows the fixed Gaussian bandwidth:

$$w_{ij} = exp\left[-\frac{1}{2}\left({d_{ij}}/{b}\right)^2\right] \qquad \textit{(6-2)}$$

where:

w_{ij} represents the geographic weight of zone j when calibrating models for zone i, d_{ij} in the case of fixed Gaussian bandwidth;

d_{ij} is the distance between the j^{th} and i^{th} nearest neighbor zones; and

b is the bandwidth.

The cross-validation technique shown in Equation (6-3) was used to determine the optimal bandwidth for the analysis described in this chapter:

$$CV = \sum_{j=1}^{n}[y_i - \hat{y}_{\neq i}(b)]^2 \qquad \textit{(6-3)}$$

where:

CV is the cross-validation error;

n is the numbers of zones; and

$\hat{y}_{\neq i}(b)$ is the adjusted value to point i omitting the observation of i.

$\hat{y}_{\neq i}(b)$ is invariably the fitted value of y_i when the jth zone is omitted during the calibration process.

The functional forms developed for Regina's total crimes and total collisions are given in Table 6-2. $\beta_{0(u_j,v_j)} \dots \dots \dots \dots \beta_{8(u_j,v_j)}$ represent the parameters obtained for each variable for each zone (u_j, v_j). j varies from zone 1 to zone 244.

Theoretically, there can be 244 different values for each parameter. Table 6-3a and 6-3b, shows some examples of the distribution of parameters in the total crimes prediction model and total collisions prediction model respectively. For example, the 50th percentile value (median) for the Proportion of Persons Age 65 plus (POP_65plus) variable is -2.782 and the range is from -2.870 (minimum) to -2.662 (maximum).

The Empirical Bayes (EB) method was used to improve the prediction accuracy of the crime and collision prediction models. The EB method is known to reduce regression-to-the-mean bias (Highway Safety Manual, 2010). Equation (6-4) shows the mathematical expression for the EB method:

$$E[\bar{K}_i] = w_g \mu_i + (1 - w_g) y_i \qquad (6\text{-}4)$$

where

$E[\bar{K}_i]$ represents the EB adjusted predicted values; and

y_i and μ_i are the observed and predicted values for location i.

w_g represents the weight (0-1) and is estimated using the global dispersion parameter (α) obtained from the GWNBR model.

Table 6-2: Functional Forms of GWNBR Models Employed in Predicting Crimes and Collisions

Categories	Functional Forms
Total Crimes	$\exp^{\beta_{0(u_j,v_j)}+\beta_{1(u_j,v_j)}\,ln(TOT\_POP_j)+\beta_{2(u_j,v_j)}POP\_65plus_j+\beta_{3(u_j,v_j)}RETAIL\_SPACE_J+\beta_{4(u_j,v_j)}OFFICE\_SPACE_J+\beta_{5(u_j,v_j)}URBAN\_HOLDING\_AREA_J+\beta_{6(u_j,v_j)}TAZ\_AREA_J}$
Total Collisions	$\exp^{\beta_{0(u_j,v_j)}+\beta_{1(u_j,v_j)}\,ln(VKMT_j)+\beta_{2(u_j,v_j)}INT\_DEN_j+\beta_{3(u_j,v_j)}NO\_3LEGS\_INT_J+\beta_{4(u_j,v_j)}NO\_LU\_PER\_TAZ_J+\beta_{5(u_j,v_j)}AVE\_SEGLEN_J+\beta_{6(u_j,v_j)}COMMERCIAL\_AREA_J+\beta_{7(u_j,v_j)}TOT\_SEGLEN\_DEN_J+\beta_{8(u_j,v_j)}LOCAL\_ROAD\_DEN_J}$

Table 6-3a: Distribution of Parameters for GWNBR Total Crime Prediction Model

Crimes Category	Variables	Min	1st Quart	Median	3rd Quart	Max
Total Crimes	Intercept	2.385	2.397	2.402	2.405	2.413
	LOG_TOT_POP	0.245	0.247	0.248	0.249	0.250
	POP_65plus	-2.870	-2.801	-2.782	-2.740	-2.662
	RETAIL_SPACE	3.439	3.468	3.481	3.495	3.537
	OFFICE_SPACE	0.490	0.509	0.519	0.525	0.545
	URBAN_HOLDING_AREA	-4.244	-4.201	-4.184	-4.172	-4.126
	TAZ_AREA	0.339	0.370	0.382	0.402	0.436
	Over-dispersion (α)			1.567		

Table 6-3b: Distribution of Parameters for GWNBR Total Collision Prediction Model

Severity	Variables	Min	1st Quart	Median	3rd Quart	Max
Total	Intercept	-6.331	-3.964	-2.923	-2.561	-2.483
	LOG_TAZ_VKMT	0.515	0.555	0.585	0.674	0.926
	INT_DEN	-0.002	0.005	0.007	0.009	0.035
	NO_3LEGS_INT	-0.029	-0.022	-0.019	-0.013	0.008
	NO_LU_PER_TAZ	-0.018	0.028	0.053	0.076	0.137
	AVE_SEGLEN	-0.094	0.030	0.042	0.052	0.097
	COMMERCIAL_AREA	0.857	0.899	1.006	1.209	1.946
	TOT_SEGLEN_DEN	-0.238	-0.098	-0.091	-0.071	-0.029
	LOCAL_ROAD_DEN	0.046	0.090	0.097	0.121	0.214
	Over-dispersion (α)			2.463		

6.9 Hotzone Identification

Regina’s crime and collision hotzones were identified by importing the number of crimes and number of collisions predicted by the models for each zone into ArcGIS. The ten 10 riskiest crime zones and ten riskiest collision zones were identified by ranking. Areas where the ten crime and ten collision zones overlapped were deemed DDACTS zones, i.e., the areas where law enforcement should be focused.

6.10 Results of Analysis

6.10.1 Crime and Collision Prediction Models

Most of the 18 independent variables listed in Table 6-1 were not statistically significant at the 95% confidence level (see Appendices A and G). Table 6-3a shows the six variables that were significant in the crime prediction model for total crimes. Table 6-3b shows the eight variables that were significant in the collision prediction model for total collisions.

The retail space and office space variables are interesting examples. In Table 6-3b, these variables have positive signs indicating that a higher proportion of retail space or office space in a zone is associated with a higher number of crimes in that zone. This result is to be expected as retail space and office space may attract, for example, break and enter criminals.

A higher proportion of commercial area in a zone is also associated with a higher number of collisions in that zone. This result is also to be expected as commercial areas are usually found in plazas or malls located on arterial roads where traffic volumes (and the number of collisions) are higher than on residential roads.

Figure 6-1 shows the Top 10 zones with the highest number of total crimes, and Figure 6-2 shows the Top 10 zones with the highest number of total collisions. The two sets of top ten zones are not identical, but both sets are located in or close to Regina's Central Business District (CBD), i.e., in areas with a high concentration of retail, office and commercial areas and a high level of traffic.

The two sets of zones identified clearly tend to be close to each other. Nine of the top ten crime zones are in the CBD or surrounding area especially the area immediately north and west of the CBD. One top ten crime zone is located to the north. Seven of the top collision zones are in the CBD or surrounding area especially the area immediately north and west of the CBD. Three top ten collision zones are located east of the CBD.

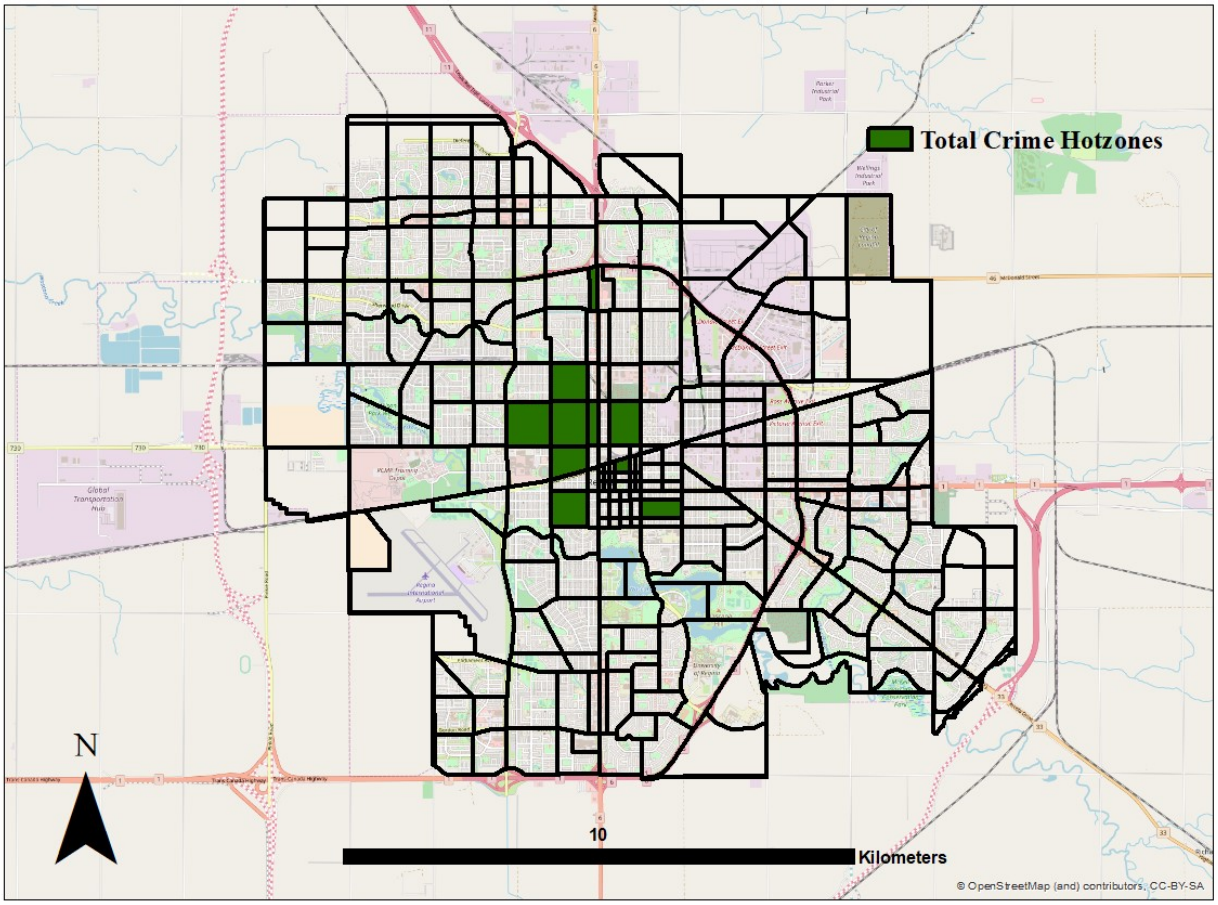


Figure 6-1: Top 10 Hotzones for Total Crimes in Regina

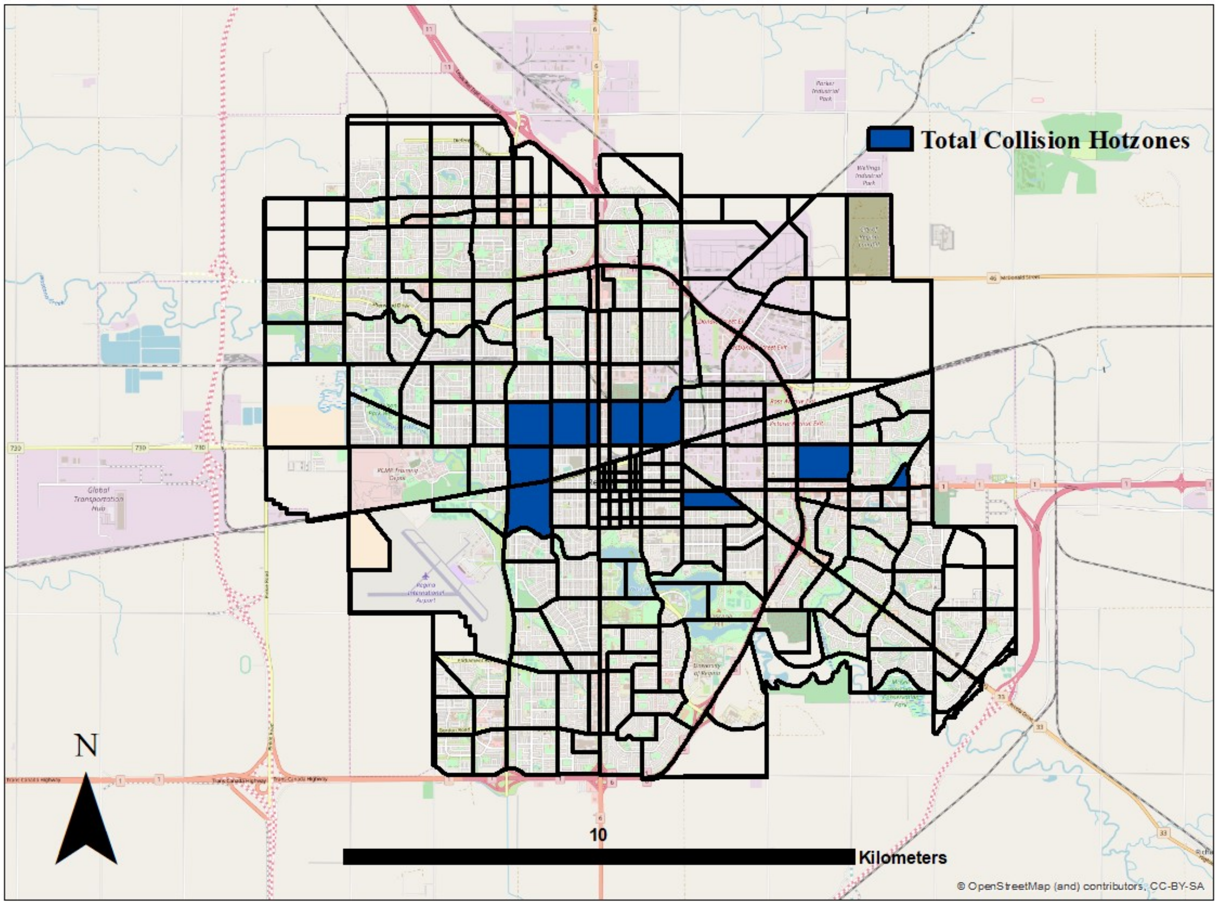

Total Collision Hotzones
N
10
Kilometers
© OpenStreetMap (and) contributors, CC-BY-SA
Global Transportation Hub

Figure 6-2: Top 10 Hotzones for Total Collisions in Regina

6.10.2 Determination and Analysis of DDACTS Zones

DDACTS zones are those that rank high for both crimes and collisions and are therefore priority areas for law enforcement. Figure 6-3 shows four zones where Regina's Top 10 crime zones and Top 10 collision zones overlap. These are the DDACTS zones.

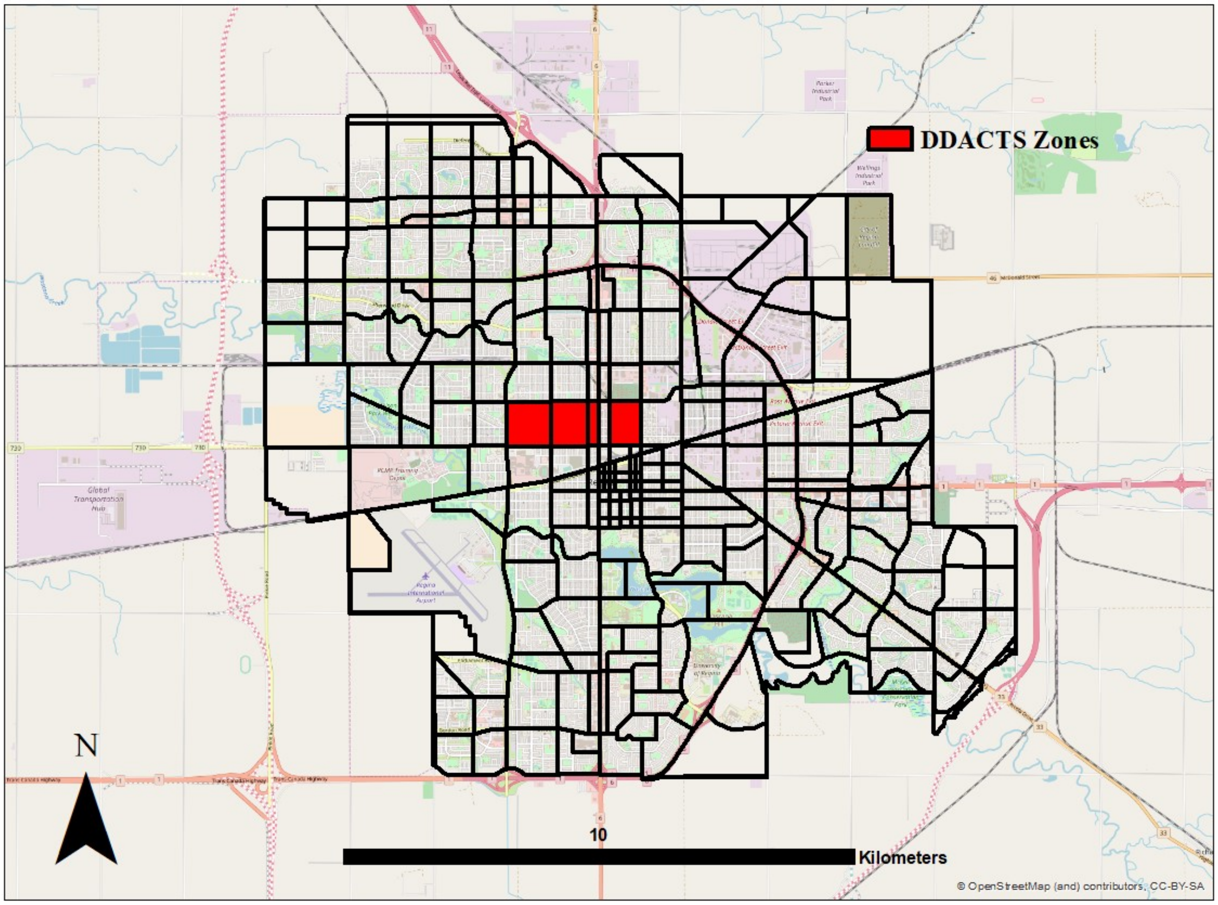

🟥 DDACTS Zones
Global Transportation Hub
Regina International Airport
N
10
Kilometers
© OpenStreetMap (and) contributors, CC-BY-SA

Figure 6-3: DDACTS Zones for Total Crimes and Total Collisions

Table 6-4 compares of the 244 TAZ zones, the top 10 hotzones and the four DDACTS Zones for Total Crimes and Total Collisions. The DDACTS hotzones combine the top ten crime hotzones with the data for the top ten collision zones. These zones represent areas where a significant numbers of crimes and collision occur or are likely to occur making these zones suitable for consideration for enforcement prioritization.

Table 6-4: Comparison between the 244 TAZ zones, the Top 10 Hotzones and the 4 DDACTS Zones for Total Crimes and Total Collisions

Categories	Total Crimes	Total Collisions
All 244 zones	50,284 crimes	26,642 collisions
Top 10 hotzones	10,633 crimes	3,546 collisions
4 DDACTS zones	5,466 crimes	1,548 collisions
Top 10 hotzones/All 244 Zones	21.1%	13.3%
DDACTS zones/ All 244 Zones	10.9%	5.8%
Area of all zones	147.93 km^2	147.93 km^2
Area of Top 10 hotzones (km^2)	4.04 km^2	5.58 km^2
Area of 4 DDACTS zones	2.11 km^2	2.11 km^2
Area of Top 10 hotzones/area of all zones	2.7%	3.8%
Area of 4 DDACTS hotzones/area of all zones	1.4%	1.4%

Important points emerge from Table 6-4:

- The City of Regina covers 147.93 km^2;
- The Top 10 crime zones cover 2.7% of the City's area, but account for 21.1% of total crimes during the study period;
- The Top 10 collision zones cover 3.8% of the City's area, but account for 13.3% of total collisions during the study period; and
- The DDACTS zones cover 1.4% of the City's area, but account for 10.9 % of total crimes and 5.8 % of the total collisions during the study period.

These findings clearly show that small areas of a city can account for a large number of both crimes and collisions (see Table 6-4). These small areas mostly have moderate to high population, but relatively poor socio-economic characteristics.

6.11 Summary, Conclusions and Future Research

The goal of this study was to identify City of Regina areas where law enforcement can be targeted to increase the effectiveness and efficiency of a city’s law enforcement by reducing the number of crimes and number of collisions simultaneously. The study had two objectives: 1) to develop macro-level crime and collision prediction models using GWNBR and the EB method; and 2) to demonstrate how the models’ results can be used to identify crime and collision hotzones where significant numbers of total crimes and total collisions are expected to overlap.

The conventional NB model cannot handle spatially correlated count data, but GWNBR models consider the distance between zones and produce different parameter values for each zone. The GWNBR approach estimates relatively similar parameters for zones that are geographically close to a target zone and less similar parameters for zones that are geographically distant. The EB method reduces regression-to-the-mean bias. This is important in improving our ability to make proactive decisions regarding future crimes and collisions. It is believed that the approach discussed here is an advance on past DDACTS studies that used GIS mapping coupled with KDE.

The study used five years (2009–2013) of data for the City of Regina, Saskatchewan. All the data were aggregated by TAZ. Two dependent variables (Total Number of Crimes and Total Number of Collisions) and the 18 independent variables were used. The explanatory variables are grouped under four headings: traffic volume, roadway, land use and socio-demographics.

The results of the study clearly show that small areas of a city can account for a large number of both crimes and collisions. The study shows that it is possible to identify the relevant areas using a data driven approach such as DDACTS. A neutral and impartial scientific procedure obviously offers advantages over an ad hoc or anecdotal approach which may appear biased and unacceptable to some citizens.

Identification of the zones where crimes and collisions overlap followed by highly visible law enforcement could help to increase the effectiveness and efficiency of employment law enforcement resources in the City of Regina and in other jurisdictions that decide to conduct a DDACTS analysis. By adopting the DDACTS approach, it should be possible to reduce the number of crimes and collisions simultaneously.

Future research should undertake a rigorous evaluation study to assess how focused law enforcement through DDACTS contributes to reducing the number of crimes and number/ severity of collisions. Future research might also consider other factors that might be important, but that were beyond the scope of this study. For example, weather conditions may have a significant influence on crime and collision occurrence. Understanding the impact of the weather and incorporating weather into the selection of DDACTS zones may help police services to target law enforcement as effectively and efficiently as possible from season to season.

References

Aguero-Valverde, J., and Jovanis, P. P. (2006). Spatial analysis of fatal and injury crashes in Pennsylvania. Accident Analysis and Prevention, 38(3), 618-625.

Anderson, J. M., MacDonald, J. M., Bluthenthal, R., and Ashwood, J. S. (2013). Reducing crime by shaping the built environment with zoning: An empirical study of Los Angeles. University of Pennsylvania Law Review. 699-756.

Andresen, M. A. (2005). Crime measures and the spatial analysis of criminal activity. British Journal of Criminology, 46(2), 258-285.

Andresen, M. A. (2007). Location quotients, ambient populations, and the spatial analysis of crime in Vancouver, Canada. Environment and Planning A, 39(10), 2423-2444.

Appiahene-Gyamfi, J. (2002). An analysis of the broad crime trends and patterns in Ghana. Journal of criminal justice, 30(3), 229-243.

Berk, R., and MacDonald, J. M (2008). Over-dispersion and Poisson regression. Journal of Quantitative Criminology, 24(3), 269-284.

Brace, Charlotte, Michelle Whelan, Belinda Clark, and Jennie Oxley (2009). The relationship between crime and road safety. (No. 284).

Castillo-Manzano, J. I., Castro-Nuño, M., and Fageda, X., (2015). Are traffic violators criminals? Searching for answers in the experiences of European countries. Transport Policy, 38, 86-94.

Cohen I. (2014). DDACTS: review of the benefits from a research perspective. 6th Edmonton's international conference on urban traffic safety, April 28-May 2, 2014 Edmonton, Alberta, Canada

Cook, C. (2012) Implementation of an area traffic officer program. A leadership white paper submitted in partial fulfillment required for graduation from the leadership command college the bill blackwood law enforcement management institute of Texas.

Cottrill, C. D., and Thakuriah, P. V. (2010). Evaluating pedestrian crashes in areas with high low-income or minority populations. Accident Analysis and Prevention, 42(6), 1718-1728.

Cui, G., Wang, X., and Kwon, D. W. A. (2015). Framework of boundary collision data aggregation into neighborhoods. Accident Analysis and Prevention, 83, 1-17.

Da Silva, A. R., and Rodrigues, T. C. V. (2014). Geographically weighted negative binomial regression—incorporating overdispersion. Statistics and Computing, 24(5), 769-783.

Dodson, E., Kirk, A., and Hill, J. (2011). Linking offence histories to accidents using OTS data: report on data collected and preliminary findings by Loughborough University.

Fell, J. C. (2013). The effects of increased traffic enforcement on other crime. In Australasian road safety research policing education conference, August 2013, Brisbane, Queensland, Australia.

Fleiter, J., Watson, A., Watson, B., and Siskind, V. (2015). Criminal histories of crash and non-crash involved Queensland speeding offenders: evidence supporting the idea that we drive as we live. Proceeding of the Australasian road safety conference, 14-16 October 2015, gold coast Australia

Frazier, A. E., Bagchi-Sen, S., Knight, J. (2013). The spatio-temporal impacts of demolition land use policy and crime in a shrinking city. Applied Geography, 41, 55-64.

Gomes, M. J. T. L., Cunto, F., and Da Silva, A. R. (2017). Geographically weighted negative binomial regression applied to zonal level safety performance models. Accident Analysis and Prevention, 106, 254-261.

Hadayeghi, A., Shalaby, A. S., and Persaud, B. N. (2010). Development of planning level transportation safety tools using geographically weighted Poisson regression. Accident Analysis and Prevention, 42(2), 676-688.

Hadayeghi, A., Shalaby, A., and Persaud, B. (2003). Macrolevel accident prediction models for evaluating safety of urban transportation systems. Transportation Research Record: Journal of the Transportation Research Board, (1840), 87-95.

Hadayeghi, A., Shalaby, A., and Persaud, B. (2007). Safety prediction models: proactive tool for safety evaluation in urban transportation planning applications. Transportation Research Record: Journal of the Transportation Research Board, 2019(1), 225-236.

Haining, R., Law, J., and Griffith, D. (2009). Modelling small area counts in the presence of overdispersion and spatial autocorrelation. Computational Statistics and Data Analysis, 53(8), 2923-2937.

Hakim, S., Shefer, D., Hakkert, A. S., and Hocherman, I. A. (1991). Critical review of macro models for road accidents. Accident Analysis and Prevention, 23(5), 379-400.

Hall, H., and Puls, E. N. (2010). Implementing DDACTS in Baltimore county: using geographic incident patterns to deploy enforcement. Geography and Public Safety, 2(3), 5-7.

Hardy, E. (2010). Data-driven policing: how geographic analysis can reduce social harm. Geography Public Safety, 2(3).

Harries, K. (2006). Extreme spatial variations in crime density in Baltimore County, MD. Geoforum, 37(3), 404-416.

Highway Safety Manual (2010). American association of state highway and transportation officials. ISBN: 9781560514770

Junger, M., West, R., and Timman, R. (2001). Crime and risky behavior in traffic: An Example of Cross-Situational Consistency. Journal of Research in Crime and Delinquency, 38(4), 439-459.

Kaygisiz, Ö, Senbil, M., and Yildiz, A. (2017). Influence of urban built environment on traffic accidents: the case of Eskisehir (Turkey). Case Studies on Transport Policy, 5, (2), 306-313.

Khondakar, B., Sayed, T., and Lovegrove, G. (2010). Transferability of community-based collision prediction models for use in road safety planning applications. Journal of Transportation Engineering, 136(10), 871-880.

Kuo, P., Lord D., and Walden T. D. (2013). Using geographical information systems to organize police patrol routes effectively by grouping hotspots of crash and crime data. Journal of Transport Geography, 30, 138-148.

Ladron de Guevara, F., Washington, S., and Oh, J. (2004). Forecasting crashes at the planning level: simultaneous negative binomial crash model applied in Tucson, Arizona. Transportation Research Record: Journal of the Transportation Research Board, (1897), 191-199.

Leigh, J. M., Dunnett J.S, and Jackson, L.M. (2016). Predictive policing using hotspot analysis. Proceedings of the International Multi-Conference of Engineers and Computer Scientists (IMECS), Vol 2, March 16 - 18, 2016, Hong Kong

Leigh, J., Dunnett, S., and Jackson, L. (2017). Predictive police patrolling to target hotspots and cover response demand. Annals of Operations Research, pp.1-16.

Lovegrove, G. R., and Sayed, T. (2006). Macro-level collision prediction models for evaluating neighbourhood traffic safety. Canadian Journal of Civil Engineering, 33(5), 609-621.

Lovegrove, G., and Sayed, T. (2017). Macrolevel collision prediction models to enhance traditional reactive road safety improvement programs. Transportation Research Record: Journal of the Transportation Research Board, (2019), 65-73.

Lovegrove, G., Lim, C., and Sayed, T. (2009). Community-based, macrolevel collision prediction model use with a regional transportation plan. Journal of Transportation Engineering, 136(2), 120-128.

Michalowski, R. J. (1975). Violence in the road: the crime of vehicular homicide. Journal of Research in Crime and Delinquency, 12(1), 30-43.

NHTSA. (National Highway Traffic Safety Administration). (2014). Data-driven approaches to crime and traffic safety, operational guidelines. Washington, DC.

Noland, R. B., and Quddus, M. A. (2003). A spatially disaggregate analysis of road casualties in England. Accident Analysis and Prevention, 36(60), 973-984.

Osgood, D. W. (2000). Poisson-based regression analysis of aggregate crime rates. Journal of Quantitative Criminology, 16(1), 21-43.

Pirdavani, A., Bellemans, T., Brijs, T., and Wets, G. (2014). Application of geographically weighted regression technique in spatial analysis of fatal and injury crashes. Journal of Transportation Engineering, 140(8), 04014032.

Piza, E. L. (2012). Using Poisson and negative binomial regression models to measure the influence of risk on crime incident counts. Rutgers Center on Public Security, 2012.

Pratt, T. (2001). Assessing the relative effects of macro-level predictors of crime: a meta-analysis. Doctoral dissertation submitted in partial fulfilment of the requirement for the award of Doctor of Philosophy, University of Cincinnati, USA.

Rhee, K. A., Kim, J. K., Lee, Y. I., and Ulfarsson, G. F. (2016). Spatial regression analysis of traffic crashes in Seoul. Accident Analysis and Prevention, 91, 190-199.

Rydberg, J., E.F. McGarrell, and A. Norris (2014). Flint DDACTS pilot evaluation. East Lansing, MI: Michigan Justice Statistics Center, School of Criminal Justice, Michigan State University.

Rydberg, J., McGarrell, E. F., Norris, A., and Circo, G. A (2018). Quasi-experimental synthetic control evaluation of a place-based police-directed patrol intervention on violent crime. Journal of Experimental Criminology, 14(1), 83-109.

Shimko G., (2013). Urban traffic safety - the city of Edmonton office of traffic safety (OTS). Canadian Injury Prevention and Safety Promotion Conference, Montreal, Quebec, November 5-7, 2013.

Siddiqui, C., Abdel-Aty, M., and Choi, K. (2012). Macroscopic spatial analysis of pedestrian and bicycle crashes. Accident Analysis and Prevention, 45, 382-391.

Silverii, S. (2010). Traffic safety initiative modernizes resource deployment in lafourche parish. Geography and Public Safety, 2(3), 7-8.

Soh, M. B. C. (2012). Crime and urbanization: revisited Malaysian case. Procedia-Social and Behavioral Sciences, 42, 291-299.

Song, J., Andresen, M. A., Brantingham, P. L., and Spicer, V. (2015). Crime on the edges: patterns of crime and land use change. Cartography and Geographic Information Science.

Stark, R. (1987). Deviant places: a theory of the ecology of crime. Criminology, 25(4), 893-910.

Statistics Canada (2016). Police-reported Crime Statistics. 2016

Statistics Canada (2017). Annual demographic estimates: subprovincial areas, July 1, 2017. Catalogue No. 91-214-X, 2018. ISSN 1920-8154

Stuster, J. (2001). Albuquerque police department: safe streets program. (Report No. DOT HS 809 278).

Takyi, E. and Park P. (2015). Identification of crime and collision hotspots for law enforcement: case study for the city of Regina. Canadian Multidisciplinary Road Safety Conference. Ottawa, 2015.

Takyi, E. Oluwajana, S.D. and Park, P.Y. (2018). Development of macro-level crime and collisions prediction models to support data-driven approach to crimes and traffic safety (DDACTS). Transport Research Record, 2018, DOI: 10.1177/0361198118777356

Taylor, R. B., Ratcliffe, J. H., and Perenzin, A. (2015). Can we predict long-term community crime problems? The estimation of ecological continuity to model risk heterogeneity. Journal of Research in Crime and Delinquency, 52(5), 635-657.

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

Troy, A., and Grove, J. M. (2008). Property values, parks, and crime: a hedonic analysis in Baltimore, MD. Landscape and Urban Planning, 87(3), 233-245.

Tseloni, A. (2006). Multilevel modelling of the number of property crimes: household and area effects. Journal of the Royal Statistical Society: Series A (Statistics in Society), 169(2), 205-233.

Walters, G. D. (2016). Proactive and reactive criminal thinking, psychological inertia, and the crime continuity conundrum. Journal of Criminal Justice, 46, 45-51.

Watson, B., Watson, A., Siskind, V., Fleiter, J., and Soole, D. (2015). Profiling high-range speeding offenders: investigating criminal history, personal characteristics, traffic offences, and crash history. Accident Analysis and Prevention, 74, 87-96.

Wedagama, D. P., Bird, R. N., and Metcalfe, A. V. (2006). The influence of urban land-use on non-motorized transport casualties. Accident Analysis and Prevention, 38(6), 1049-1057.

Wegman, F. (2017). The future of road safety: a worldwide perspective. IATSS Research, *40*(2), 66-71.

Wiebe, D. J., Ray, S., Maswabi, T., Kgathi, C., and Branas, C. C. (2016). Economic development and road traffic fatalities in two neighbouring African nations. African Journal of Emergency Medicine, Vol. *6*, No. 2, pp. 80-86.

Wier, M., Weintraub, J., Humphreys, E. H., Seto, E., and Bhatia, R. (2009). An area-level model of vehicle-pedestrian injury collisions with implications for land use and transportation planning. Accident Analysis and Prevention, 41(1), 137-145.

Wisconsin Department of Transportation (Wisconsin DOT) (2010). DDACTS: targeting crash and crime hotspot in Wisconsin traffic safety reporter, 13(3), 1-3.

Wolfe, M. K., and Mennis, J. (2012). Does vegetation encourage or suppress urban crime? Evidence from Philadelphia, PA. Landscape and Urban Planning, 108(2), 112-122.

Wyatt, J., and Alexander, M (2010). Integrating crime and traffic crash data in Nashville. Geography and Public Safety, 2(3).

Yu, C. Y., and Xu, M. (2017). Local Variations in the Impacts of Built Environments on Traffic Safety. Journal of Planning Education and Research, 0739456X17696035.

Yu, D. (2010). Exploring spatiotemporally varying regressed relationships: the geographically weighted panel regression analysis. In Proceedings of the Joint International Conference on Theory, Data Handling and Modelling in GeoSpatial Information Science, 134-139.

CHAPTER 7:CONCLUSION AND RECOMMENDATION FOR FUTURE RESEARCH

This chapter starts with a summary and set of conclusions from the studies discussed in Chapters 4 to 6. It then presents the main contributions of this study and the main recommendations for future research.

7.1 Summary and Conclusion

This dissertation consisted of three main technical components.

The first component (Chapter 4) explored the development of collision prediction models using two approaches to spatial regression analysis, i.e., the two commonly used spatial models: GWPR and GWNBR. These models were used to account for spatial dependency in the development of collision prediction models using one of two bandwidth types, fixed Gaussian bandwidth or adaptive bi-square bandwidth. The predictive performance of these models was compared, and it was concluded (from the CURE plots) that the fixed Gaussian bandwidth is a better choice for Regina's collision database.

The second component (Chapter 5) discussed crime prediction models using GWPR and GWNBR. The research focused mainly on the fixed Gaussian bandwidth in the model calibration. More importantly, it demonstrated how the EB technique can enhance identification and prediction of areas concern for security due to a high number of violent and non-violent crimes.

The third component (Chapter 6) focused on the identification of DDACTS zones. This exercise used a set of prediction models for the total number of collisions and total number of crimes.

The first component has contributed to zonal investigation by developing macro-level collision prediction models using relatively new statistical methods designed to handle the problem of spatial dependency and over-dispersion in macro-level collision data. The need for a geographically weighted model became clear since Moran's I local indicator test showed that the study data contained statistically significant levels of spatial dependency. Bandwidth is a required and important input for geographically weighted regression models. Both the fixed Gaussian bandwidth and adaptive bi-square bandwidth were tested. Unlike many previous studies which advocated adaptive bandwidth, the fixed Gaussian bandwidth was a more appropriate bandwidth in terms of the fitting performance achieved in the macro-level prediction of collisions using Regina data. This finding was supported largely by the CURE plots developed for each collision prediction model. The good fitting performance of the fixed Gaussian bandwidth could be linked to the use of consistent scale of analysis while examining the relationships across data aggregated into irregularly shaped areal units. This study also applied the seven goodness-of-fit (GOF) tests most widely used in transportation engineering (i.e., Akaike information criterion, Corrected Akaike information criterion, Bayesian information criteria, mean square error, mean square prediction error, mean prediction bias and mean absolute deviation). The results of these GOF tests were inconsistent in terms of selecting the best fitting model. This finding is in line with previous studies (Lord and Park, 2008; Young and Park, 2013).

This study concluded that it is difficult to choose a particular bandwidth method as the best bandwidth method for different study databases from different areas. The appropriate type of bandwidth method could be different from one database to another. This finding suggests that researchers need to apply both bandwidth methods before selecting the more appropriate method suitable for their particular study database.

GWPR and GWNBR produced a wide range of parameter values across zones regardless of the type of bandwidth applied. It appeared that this was due to unobserved heterogeneity issues in each zone. Although, this research used the cross-validation technique to select the optimal bandwidth, the variation in the parameters of these models could also be influenced by the method used to determine the optimal bandwidth. Other methods for choosing optimal bandwidth include the use of Akaike Information Criteria and Deviance Information Criteria.

The second component of this dissertation focused on crime prediction. The same approach discussed in the first component was used for modelling the Regina crime dataset. The EB technique was applied to mitigate the issue of RTM bias and seven GOF tests were applied. Like the collision prediction model findings, the test results were not consistent in terms of choosing the best fitting crime prediction model. This result again suggested that there is no single GOF test that will provide a concrete answer to the choice of the best fitting crime prediction model. The GWNBR model was selected to model Regina's crime data because the method allows the use of the EB technique to reduce RTM bias in crime data. As a result, the hot zones identified can be viewed as areas with long-term security concerns where police enforcement should be focused.

The last component in this dissertation used the concept of DDACTS to identify hot zones where focused law enforcement can reduce the number of crimes and the number of collisions simultaneously. This study identified four DDACTS zones in Regina. These four zones cover only 1.4% of the City's area, but account for 10.9% of expected total crimes and 5.8% of expected total collisions. The approach offers an impartial scientific procedure that identifies areas for targeted law enforcement and opportunities to increase the effectiveness and efficiency of limited law enforcement resources.

7.2 Contribution

The four DDACTS hotzones identified by this study's analyses can be regarded as contributing to a strengthened rationale for the proactive deployment of law enforcement.

This dissertation makes a major contribution to the advancement of existing DDACTS techniques in identifying zones for focused police enforcement. The dissertation proposes an alternative approach that allows incorporation of predictors of crimes and collisions rather than relying largely on kernel density estimation for hotzone determination. The proposed macro-level collision analysis has long been used by transportation professionals in long-range safety planning analysis.

As the concept of DDACTS relies on highly visible enforcement, the results of this study can possibly be used in the selection of appropriate locations for the installation of close circuit television (CCTV), with clearly visible signage stating that surveillance is being carried out, to enhance security in an area. The use of CCTV may broaden the value of DDACTS as not only an operational level enforcement tactic, but also a planning level enforcement tactic.

Over the years, advances in statistical modelling have led to the development of models that could accommodate spatial dependency, for example, the geographically weighted regression model. However, understanding the limitations of geographically weighted regression (and the assumption of normality on the response) leads to various extensions. The initial extension of the geographically weighted regression is known as the geographically weighted Poisson regression (GWPR). However, a major limitation of the GWPR is the assumption of equal dispersion, i.e., that the mean and variance are equal. Despite this limitation, the GWPR has been used extensively due to its ability to take care of the issue of spatial dependence.

Recently, the geographically weighted negative binomial regression (GWNBR), an extension that allows for negative binomial distribution and relaxes the assumption of equal dispersion, was proposed. There has not been a data-driven comparison of the two types of geographically weighted regression in the context of collision and crime prediction modelling. This is another research gap which this study has filled by rigorously testing the fitting performance of GWPR and GWNBR. The advantage of GWNBR for addressing spatial dependency and over-dispersion is emphasised in this study.

The study also evaluated the impact of different bandwidth choices on the predictive performance of geographically weighted models. Comparisons of two bandwidth types (fixed Gaussian and Adaptive bi-square) provided additional insights into the need for testing the appropriate type of bandwidth when researchers are developing geographically weighted regression models.

This study also confirmed that there is potential danger in using a single goodness-of-fit (GOF) measure in assessing the performance of collision and crime prediction models. The study shows that it may be advantageous to use CURE plots rather than GOF tests in this regard.

7.3 Future Work

While this research offers an extension to the traditional DDACTS approach and facilitates identification of hot zones at the areal unit level, there are still large opportunities for a future research.

Firstly, this research only considered fixed dispersion parameters for the GWNBR model. Comparing GWNBR with varying dispersion parameter to understand the impact on fitting performance can be viewed as future research.

Secondly, this research observed that some variables show coefficient reversal (i.e., signs change from plus to minus or vice versa) across different zones. Future research could investigate in detail to understand the reason for coefficient reversal in GWPR and GWNBR.

Thirdly, this research used the VKMT (the exposure variable) in the CURE plots to assess the fitting performance of the collision models developed. Future research could evaluate CURE plots considering other input variables used in the development of collision models. This will provide insight into how well the other independent variables performed in fitting the dependent variable.

Fourthly, transportation engineers estimate future events (crimes or collisions) by developing scenarios. In reality, city officials have growth targets that determine changes and development that influence future demographic and land use variables. Assumptions regarding anticipated changes could further contribute to the problem of RTM. While this research argued for RTM to always be considered, future work could consider the pattern of RTM in historical crime and collision data aggregated into area units. This could provide further justification for the need for RTM to be considered in macro-level modelling.

Lastly, the Modifiable Areal Unit Problem (MAUP) needs to be evaluated possibly using different types of zone (e.g., census areas instead of the traffic analysis zones used in this study) to understand more clearly the stability of the hotzones as identified in this dissertation.

References

Lord, D., and Park, P. Y. J., (2008). Investigating the effects of the fixed and varying dispersion parameters of Poisson-gamma models on empirical Bayes estimates. Accident Analysis & and Prevention, 40(4), 1441-1457.

Young, J., and Park, P.Y., (2013). Benefits of small municipalities using jurisdiction-specific safety performance functions rather than the highway safety manual's calibrated or uncalibrated safety performance functions. Canadian Journal of Civil Engineering, 40 (6), 517-527.

APPENDICES

Appendix A: Negative Binomial (NB) Model for Total, Fatal Injury and Property Damage Only Collisions

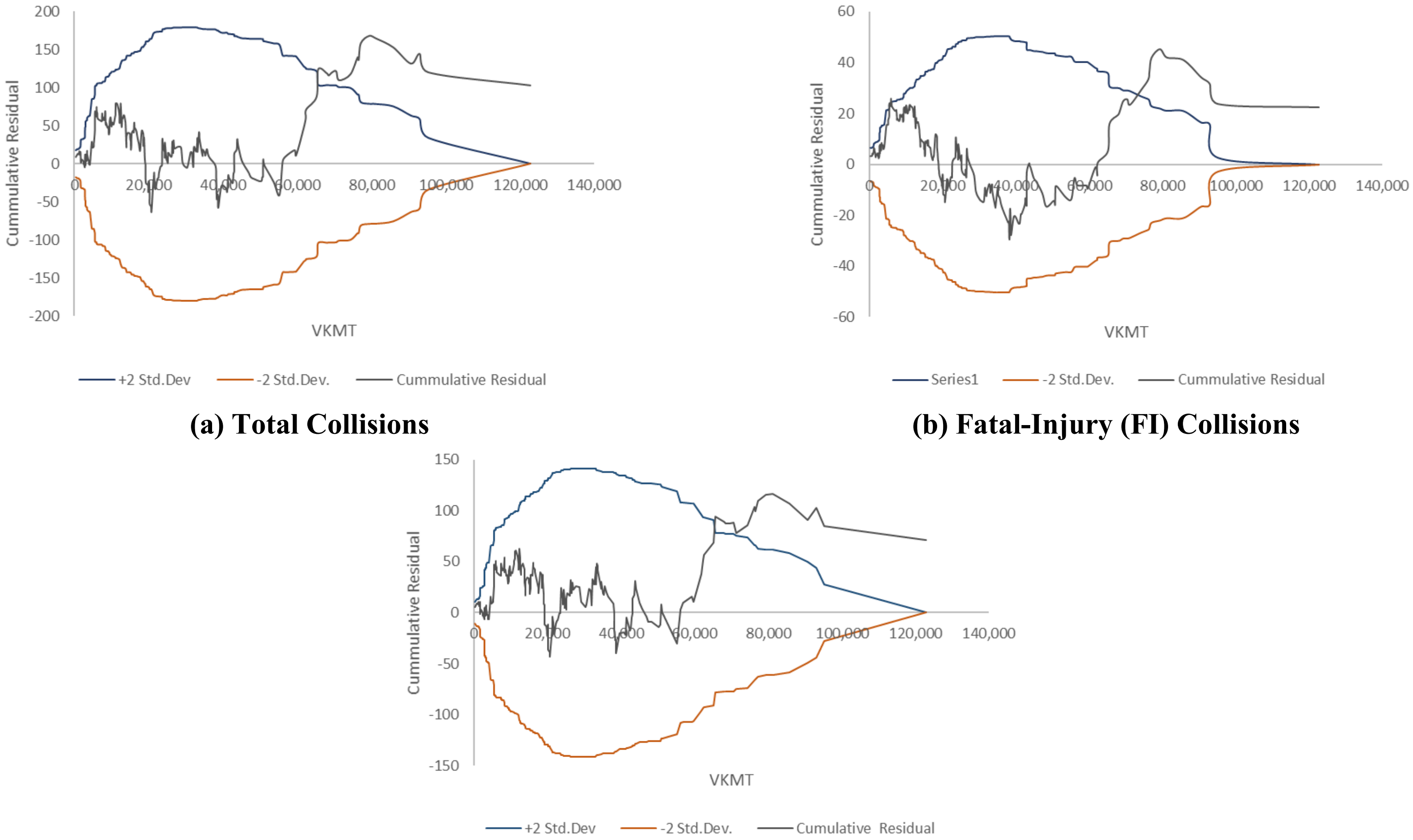
Table A-1: Functional Forms for Best Predicting Models for Total, Fatal – Injury (FI) and Property Damage Only (PDO) Collisions

Severity	Functional Form
Total	$\mu_i = exp^{\beta_0+\beta_1 \ln(VKMT)+\beta_2\times INT\_DEN+\beta_3\times NO\_3LEGS\_INT+\beta_4\times NO\_LU\_PER\_TAZ\ +\beta_5\times AVE\_SEGLEN+\beta_6\times COMMERCIAL\ \_AREA+\beta_7\times TOT\_SEGLEN\_DEN+\beta_8\times LOCAL\_ROAD\_DEN}$
FI	$\mu_i = exp^{\beta_0+\beta_1 \ln(VKMT)+\beta_2\times NO\_LU\_PER\_TAZ\ +\beta_3\times\ COMMERCIAL\ \_AREA+\beta_4\times ARTERIAL\_LEN+\beta_5\times\ RESIDENTIAL\_MD\_AREA+\beta_6\times INDUSTRIAL\_AREA}$
PDO	$\mu_i = exp^{\beta_0+\beta_1 \ln(VKMT)+\beta_2\times INT\_DEN+\beta_3\times NO\_3LEGS\_INT+\beta_4\times NO\_LU\_PER\_TAZ\ +\beta_5\times AVE\_SEGLEN+\beta_6\times COMMERCIAL\ \_AREA+\beta_7\times TOT\_SEGLEN\_DEN+\beta_8\times LOCAL\_ROAD\_DEN}$

Table A-2: Model Parameters for Total, Fatal Injury and Property Damage Only Collisions

Parameter	Total			Fatal-injury			Property Damage Only		
	Estimate	Standard Error	Pr > ChiSq	Estimate	Standard Error	Pr > ChiSq	Estimate	Standard Error	Pr > ChiSq
Intercept	-3.8105	0.4724	<.0001	-5.3251	0.5853	<.0001	-3.8295	0.4762	<.0001
LOG_TAZ_VKMT	0.6379	0.0530	<.0001	0.5859	0.0629	<.0001	0.6156	0.0533	<.0001
INT_DEN	0.0102	0.0038	0.0068	-	-	-	0.0099	0.0038	0.0093
NO_3LEGS_INT	-0.0181	0.0068	0.0078	-	-	-	-0.0176	0.0067	0.0089
NO_LU_PER_TAZ	0.0537	0.0336	0.1099	0.1362	0.0397	0.0006	0.0537	0.0333	0.1070
AVE_SEGLEN	0.0449	0.0390	0.2494	-	-	-	0.0525	0.0387	0.1749
COMMERCIAL_AREA	0.9687	0.1820	<.0001	1.0866	0.1761	<.0001	0.9826	0.1825	<.0001
TOT_SEGLEN_DEN	-0.0802	0.0208	0.0001	-	-	-	-0.0837	0.0216	0.0001
LOCAL_ROAD_DEN	0.0997	0.0196	<.0001	-	-	-	0.1036	0.0199	<.0001
ARTERIAL_LEN	-	-	-	0.1249	0.1028	0.2244	-	-	-
RESIDENTIAL_MD_AREA	-	-	-	1.0810	0.3926	0.0059	-	-	-
INDUSTRIAL_AREA	-	-	-	0.4139	0.1947	0.0335	-	-	-
Dispersion		0.3545			0.4294			0.3446	

Appendix B: CURE Plots for Negative Binomial (NB) Models



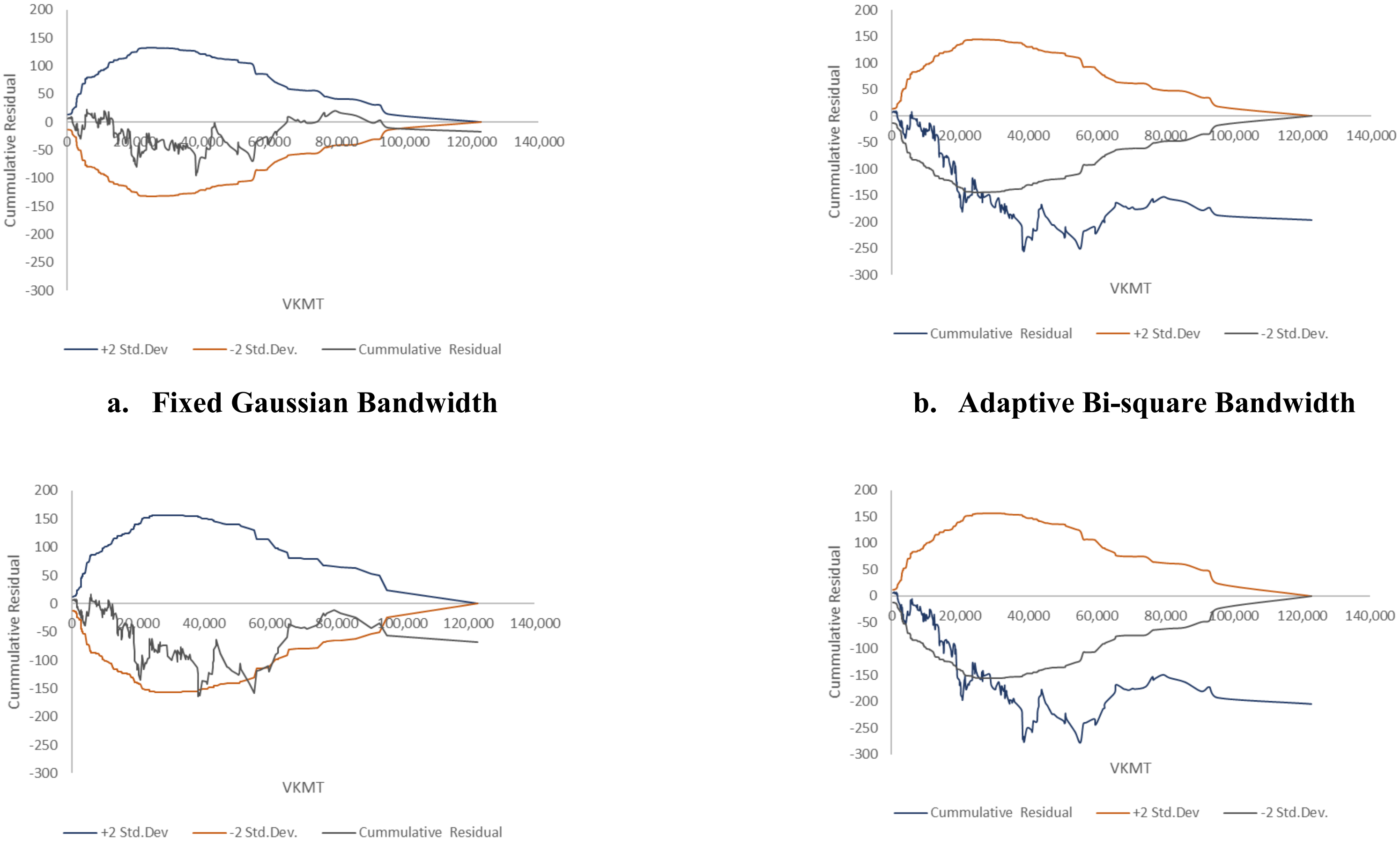
(a) Total Collisions

(b) Fatal-Injury (FI) Collisions

(c) Property Damage Only (PDO) Collisions

Figure B-1: Cumulative Residual Plot for Collisions Severity Types using Negative Binomial Model

Appendix C: CURE Plots for GWPR and GWNBR Models for Total Collisions



a. **Fixed Gaussian Bandwidth**

b. **Adaptive Bi-square Bandwidth**

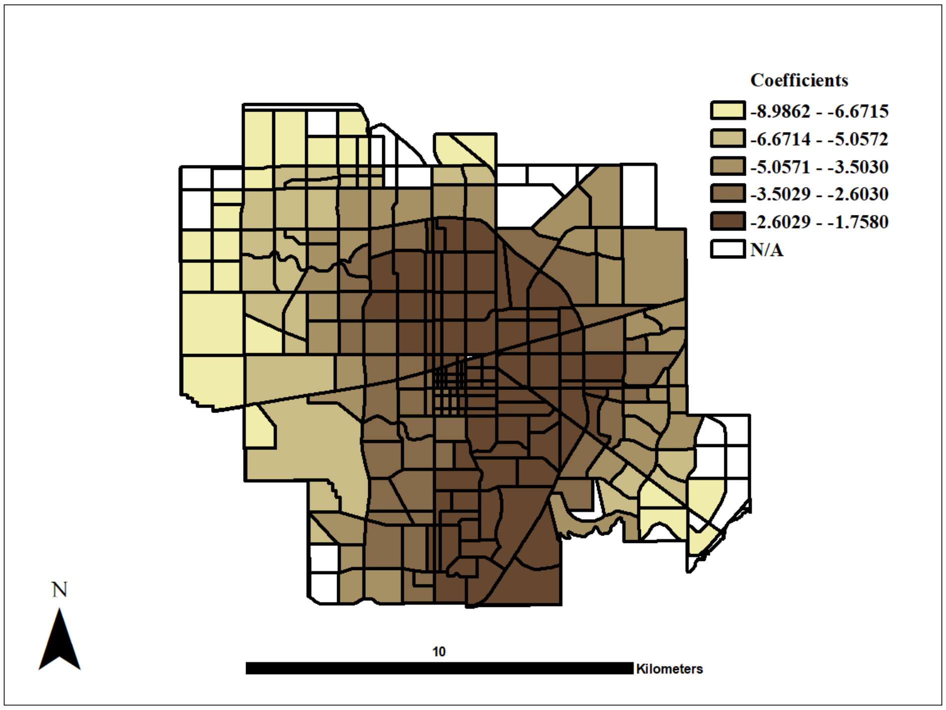
c. **Fixed Gaussian Bandwidth**

d. **Adaptive Bi-square Bandwidth**

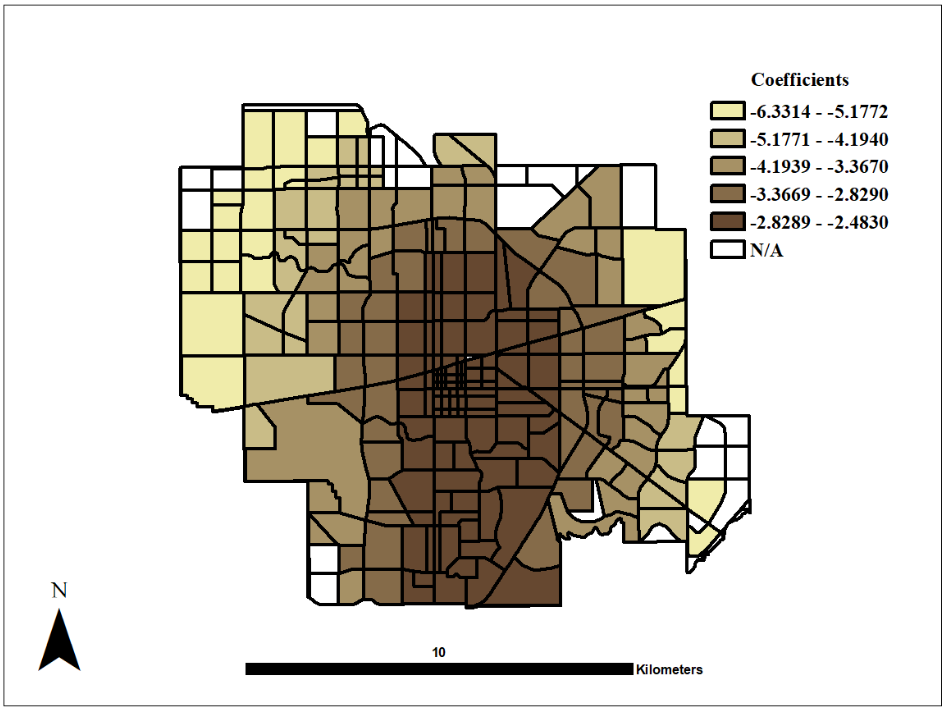
Figure C-1: CURE Plots for GWPR (a and b) and GWNBR (c and d) Showing Fixed and Adaptive Bandwidth Results

Appendix D: Thematic Representation of the Coefficients of GWPR and GWNBR Models for Total Collisions

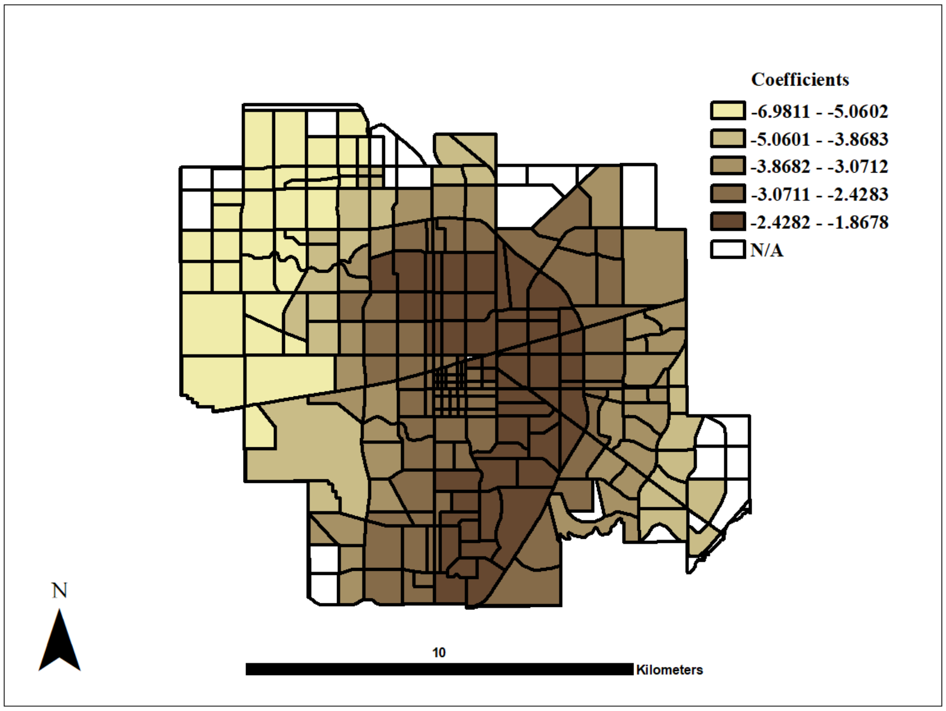
A. INTERCEPT



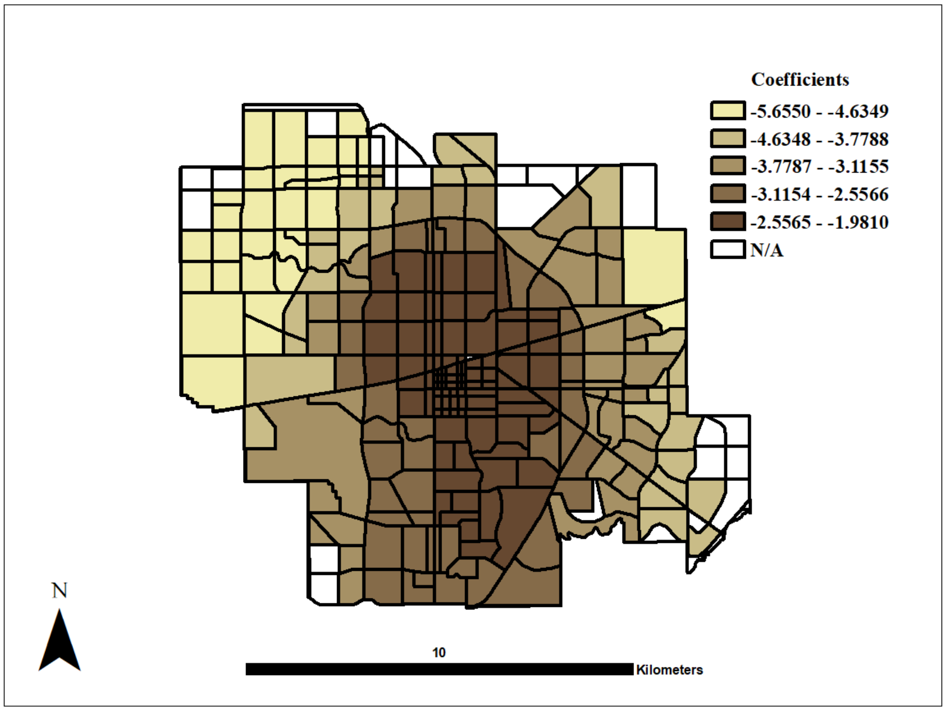
a) Fixed Gaussian Bandwidth GWPR



b) Fixed Gaussian Bandwidth GWNBR

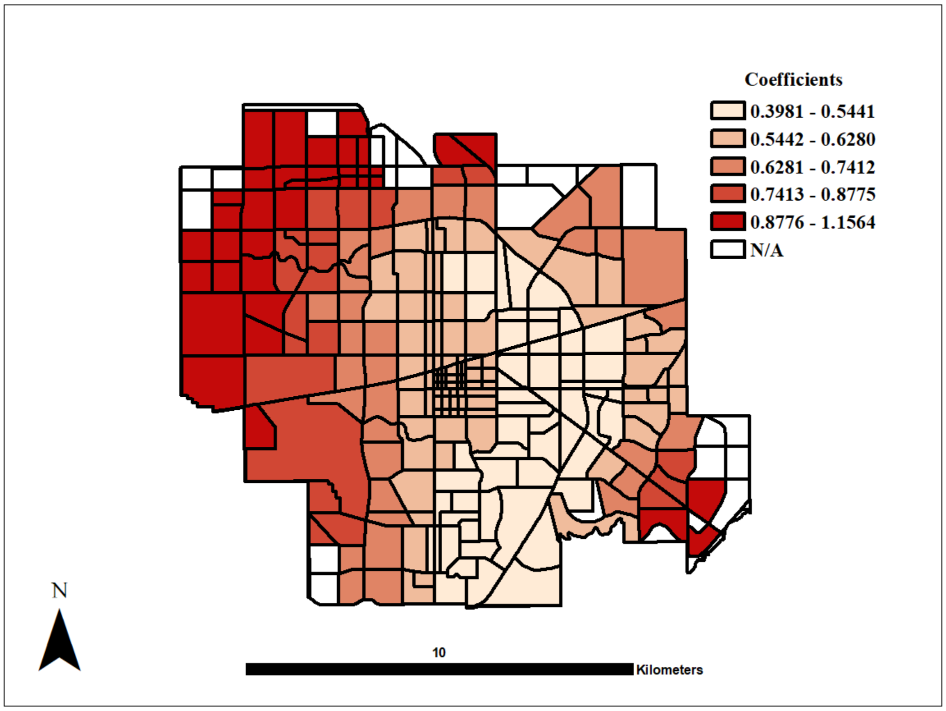


c) **Adaptive Bi-square Bandwidth GWPR**

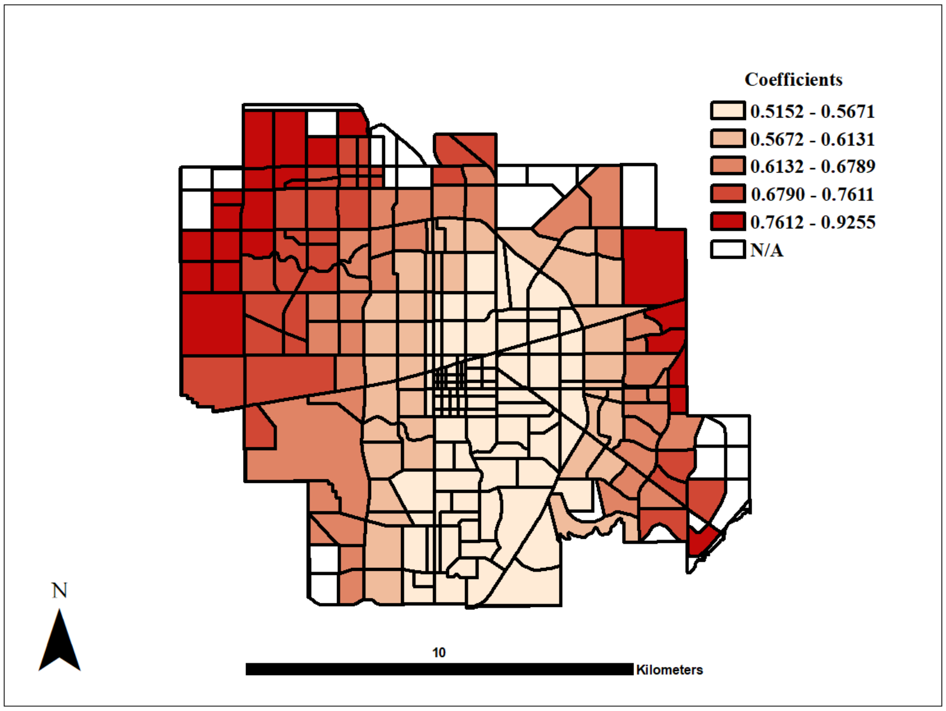


d) **Adaptive Bi-square Bandwidth GWNBR**

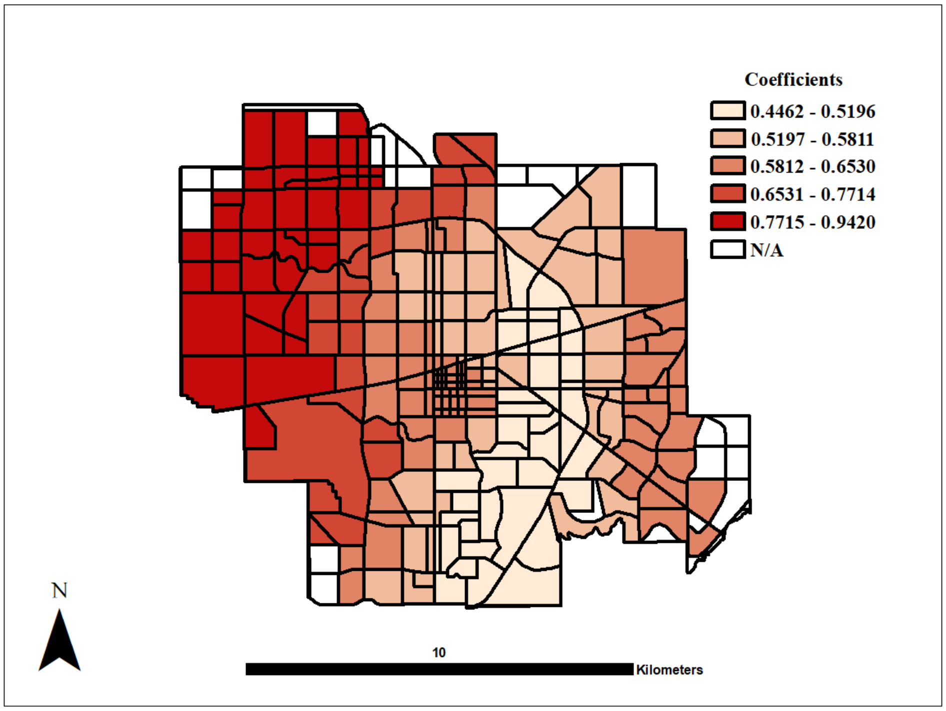
B. LOG OF TAZ VKMT



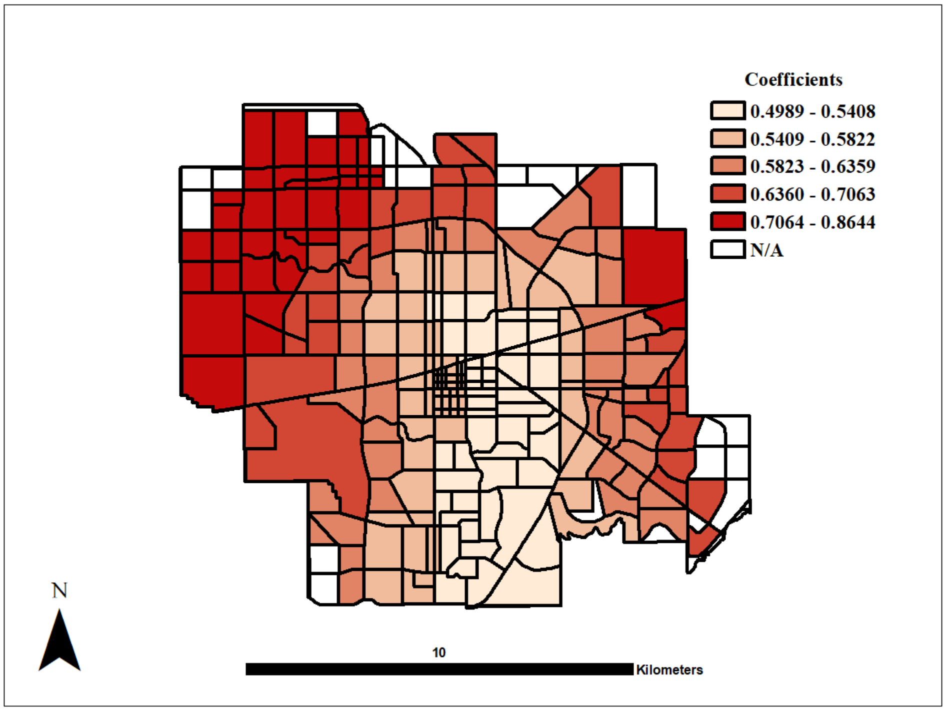
a) **Fixed Gaussian Bandwidth GWPR**



b) **Fixed Gaussian Bandwidth GWNBR**

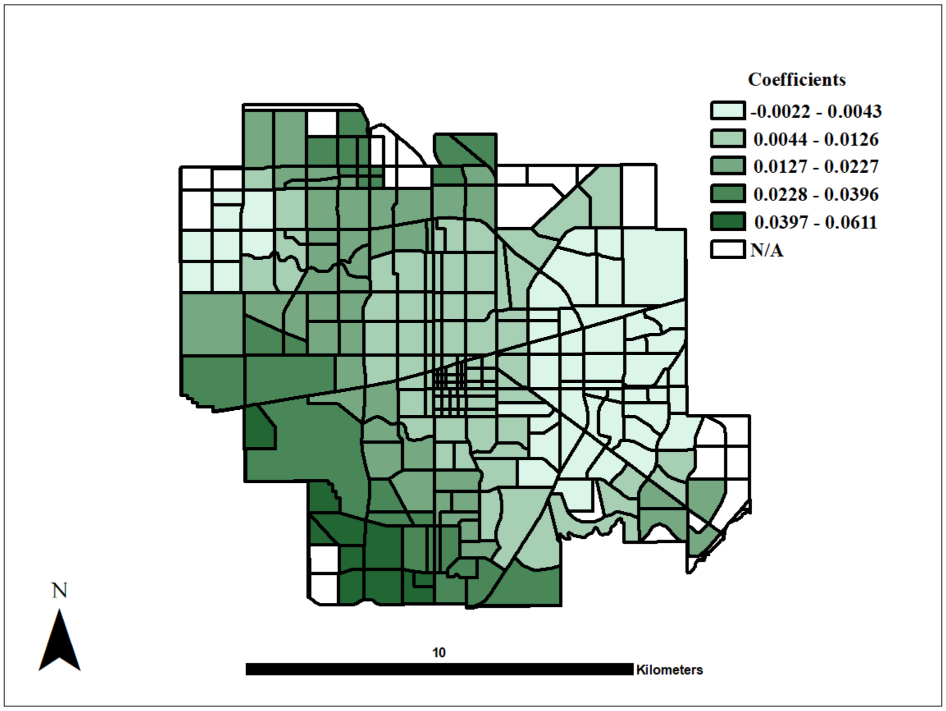


c) Adaptive Bi-square Bandwidth GWPR

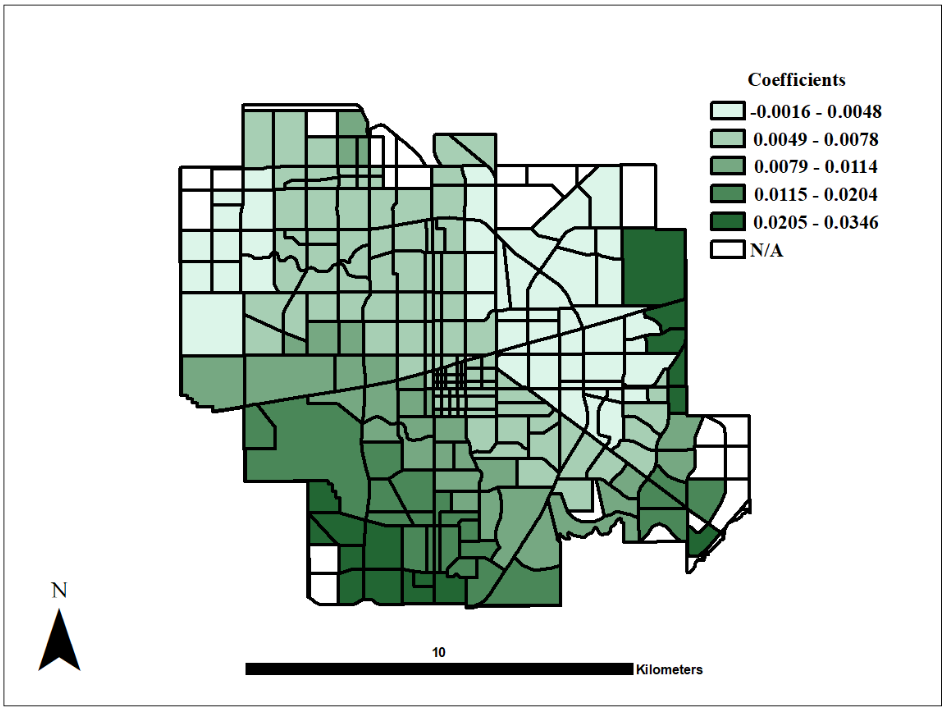


d) Adaptive Bi-square Bandwidth GWNBR

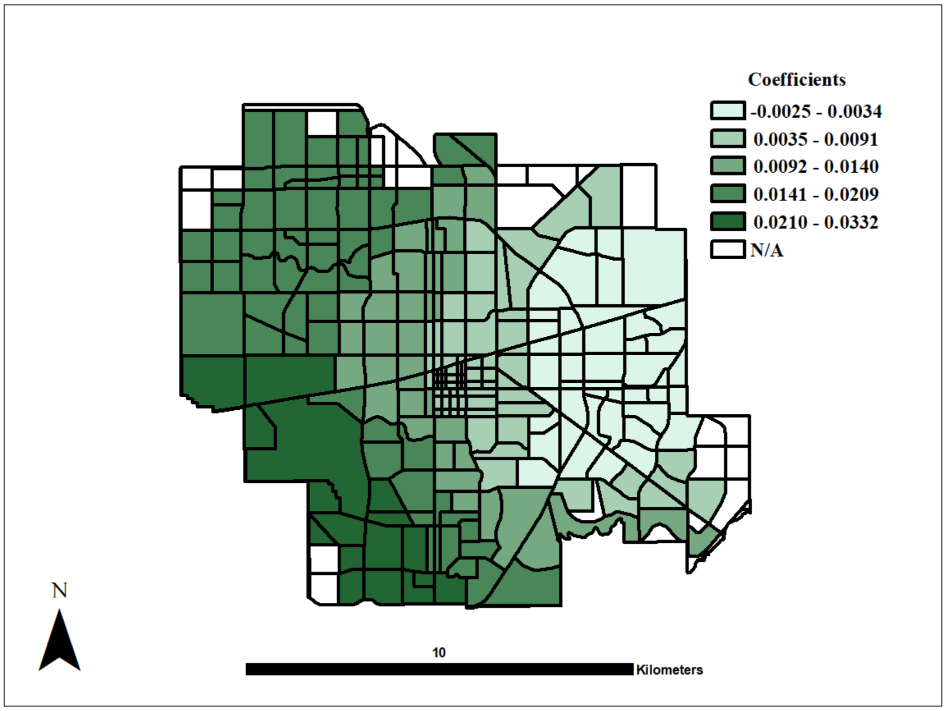
C. INTERSECTION DENSITY



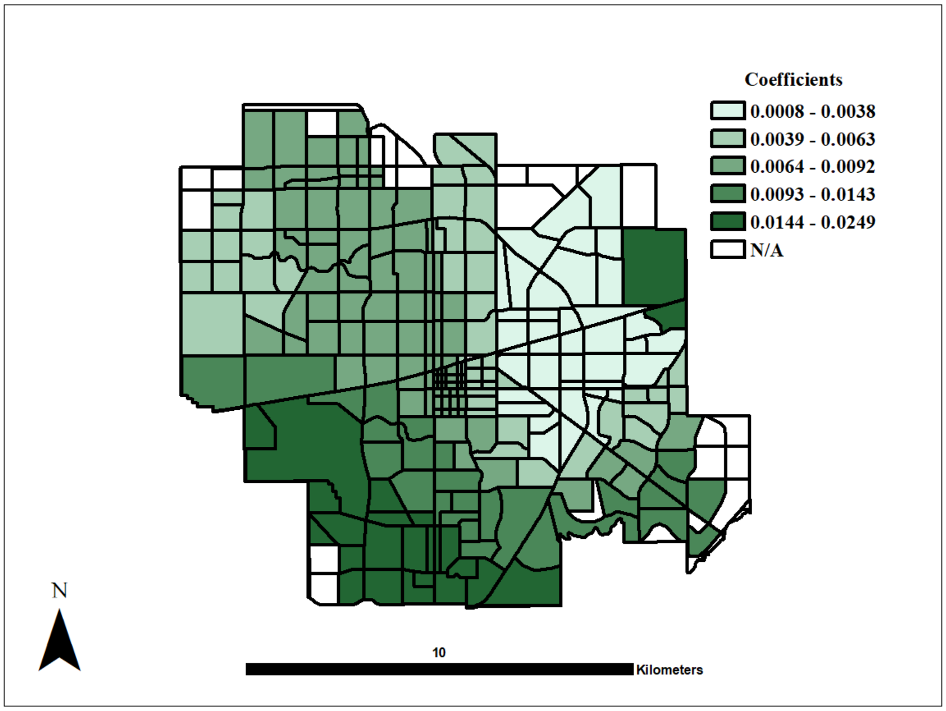
a) Fixed Gaussian Bandwidth GWPR



b) Fixed Gaussian Bandwidth GWNBR

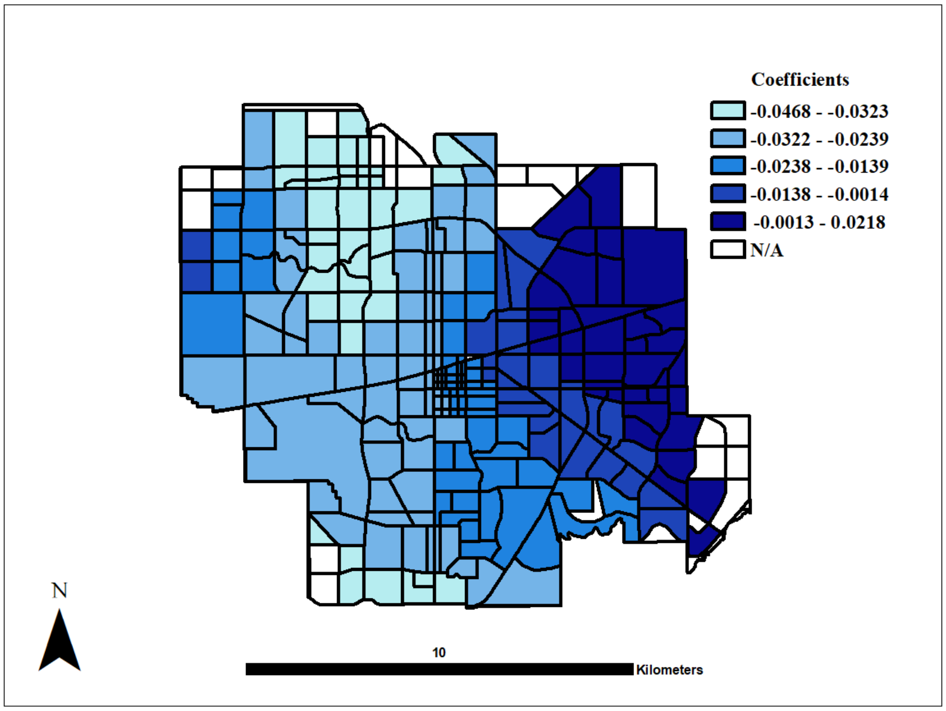


c) **Adaptive Bi-square Bandwidth GWPR**

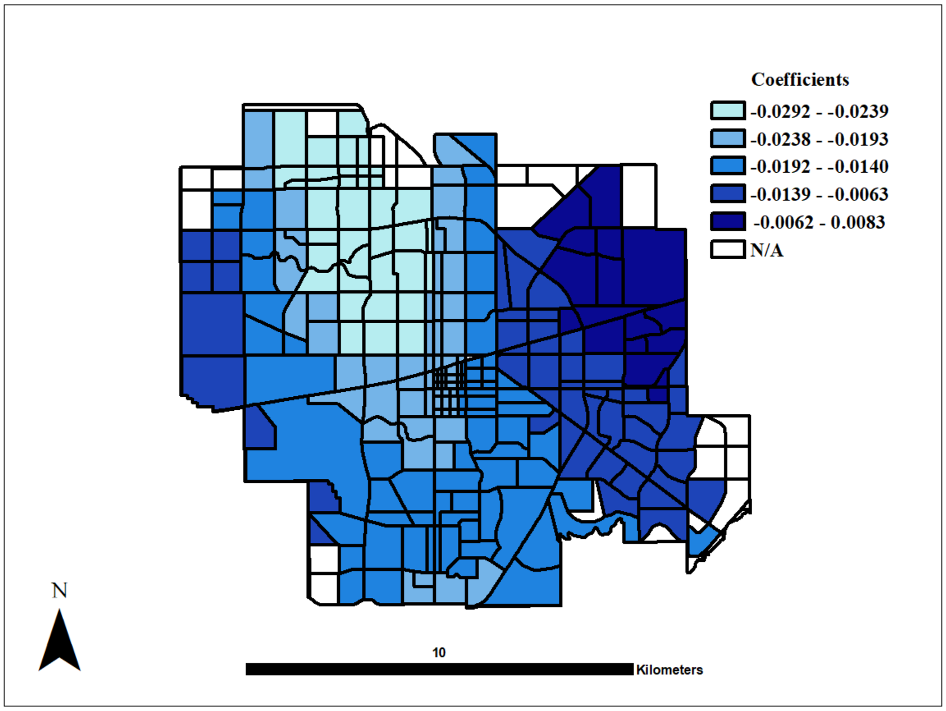


d) **Adaptive Bi-square Bandwidth GWNBR**

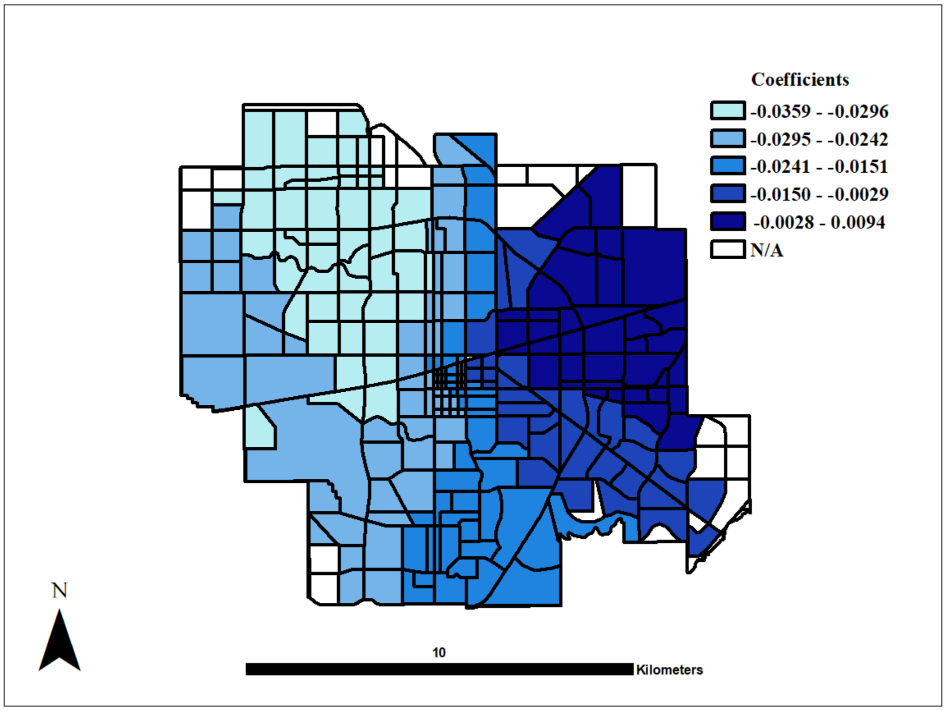
D. NO OF 3 LEG INTERSECTIONS



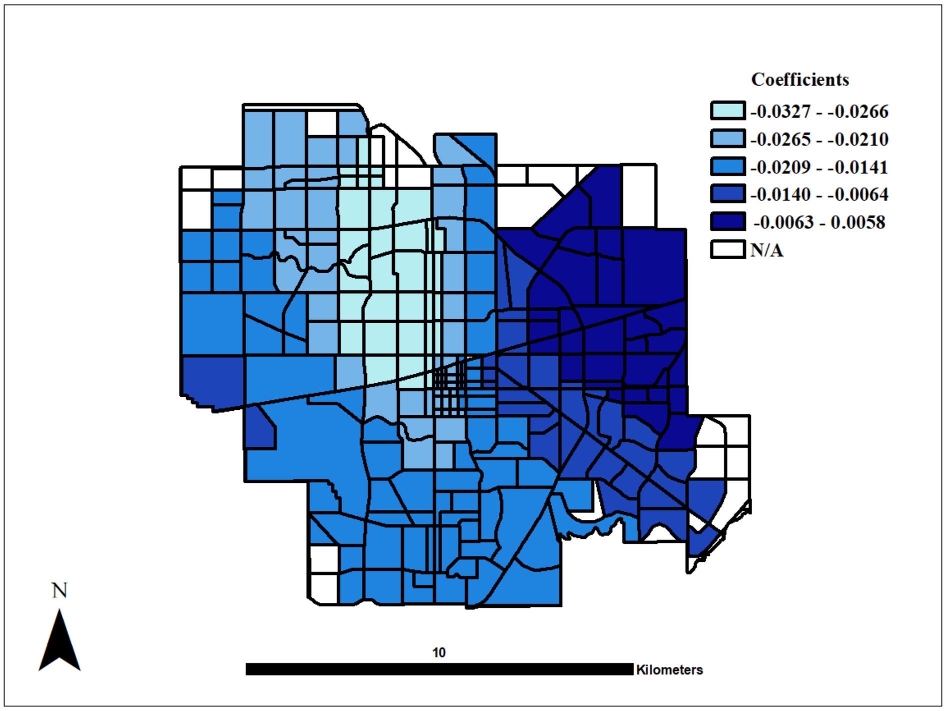
a) **Fixed Gaussian Bandwidth GWPR**



b) **Fixed Gaussian Bandwidth GWNBR**

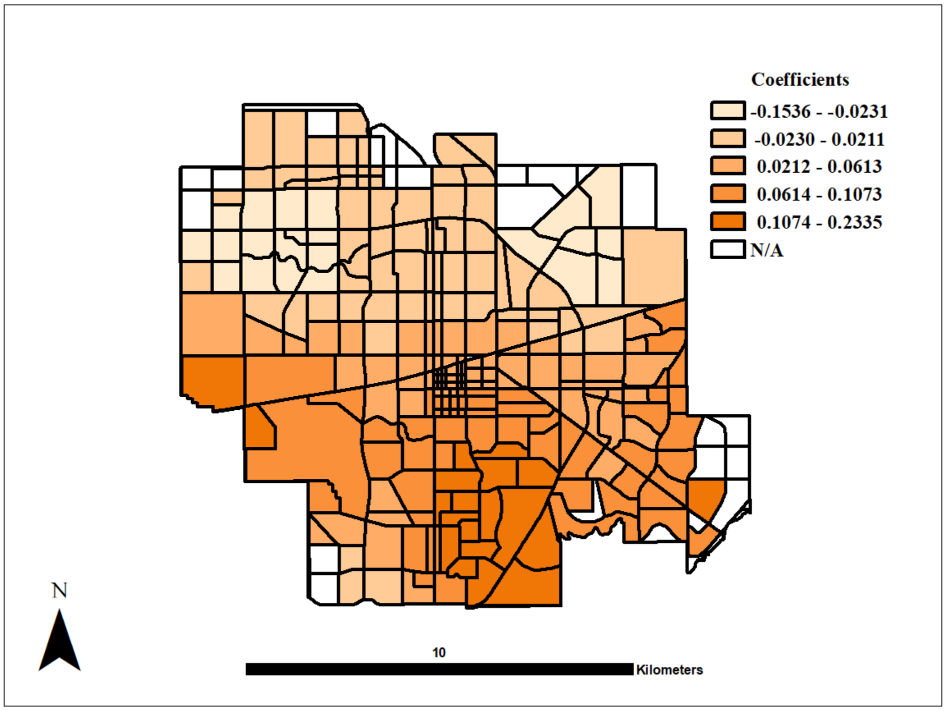


c) **Adaptive Bi-square Bandwidth GWPR**

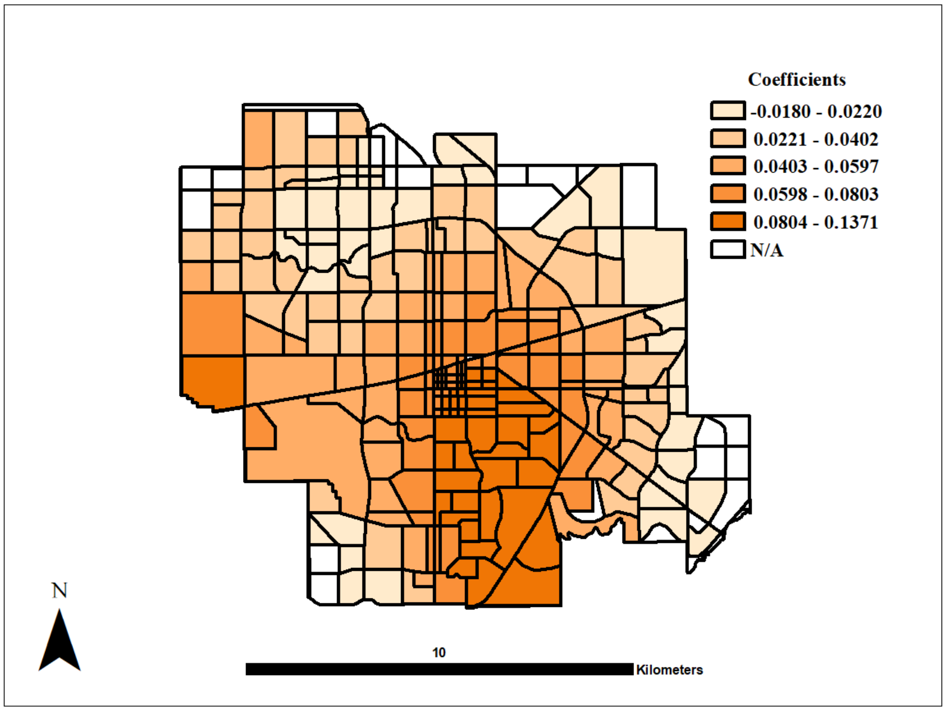


d) **Adaptive Bi-square Bandwidth GWNBR**

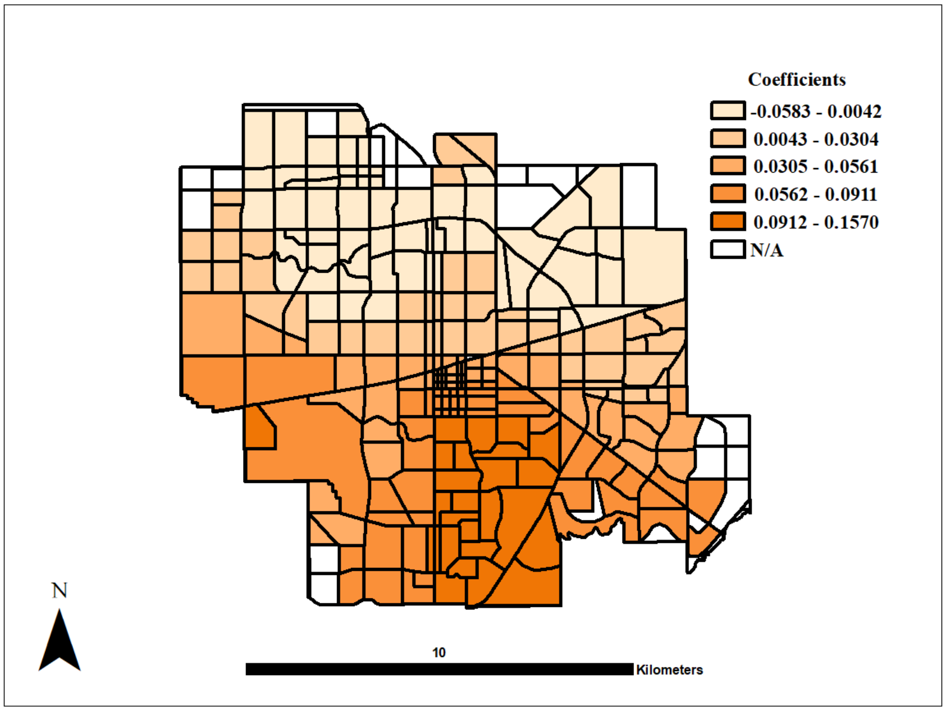
E. NUMBER OF LAND USES



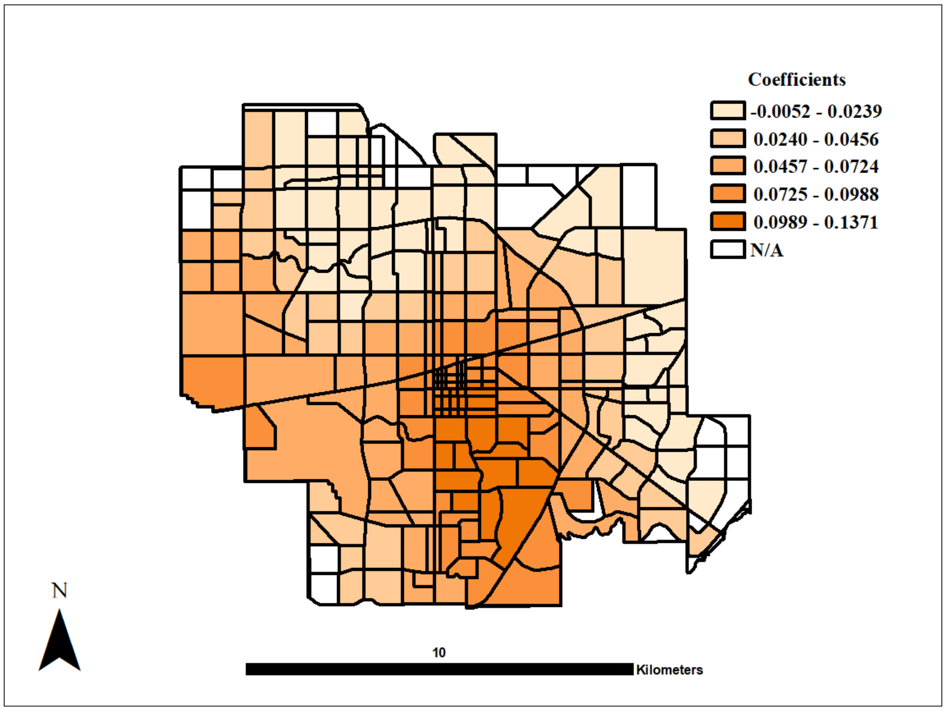
a) Fixed Gaussian Bandwidth GWPR



b) Fixed Gaussian Bandwidth GWNBR

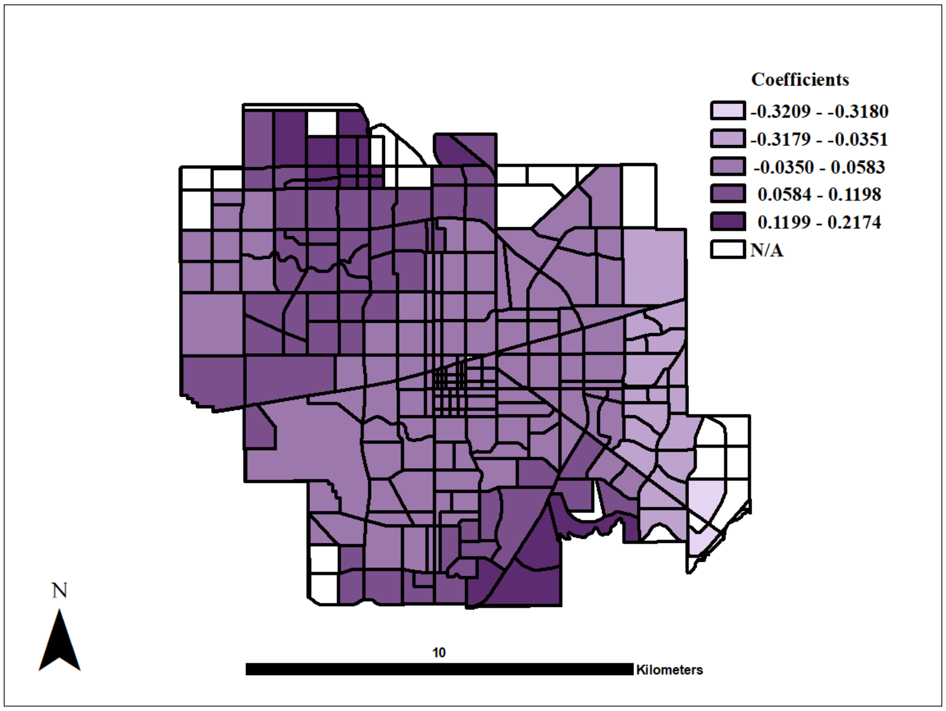


c) Adaptive Bi-square Bandwidth GWPR

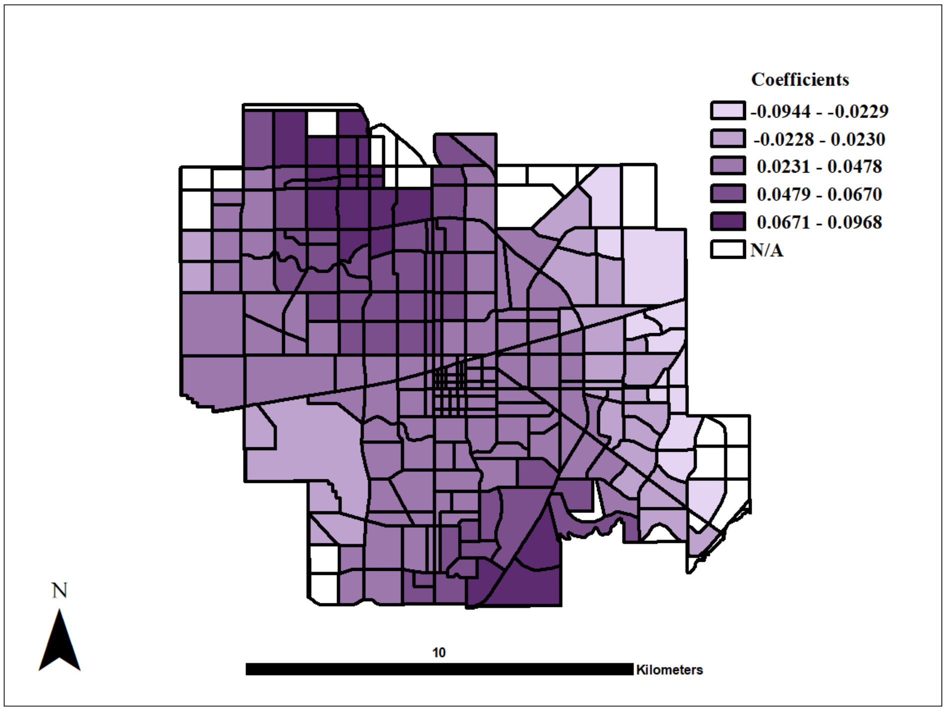


d) Adaptive Bi-square Bandwidth GWNBR

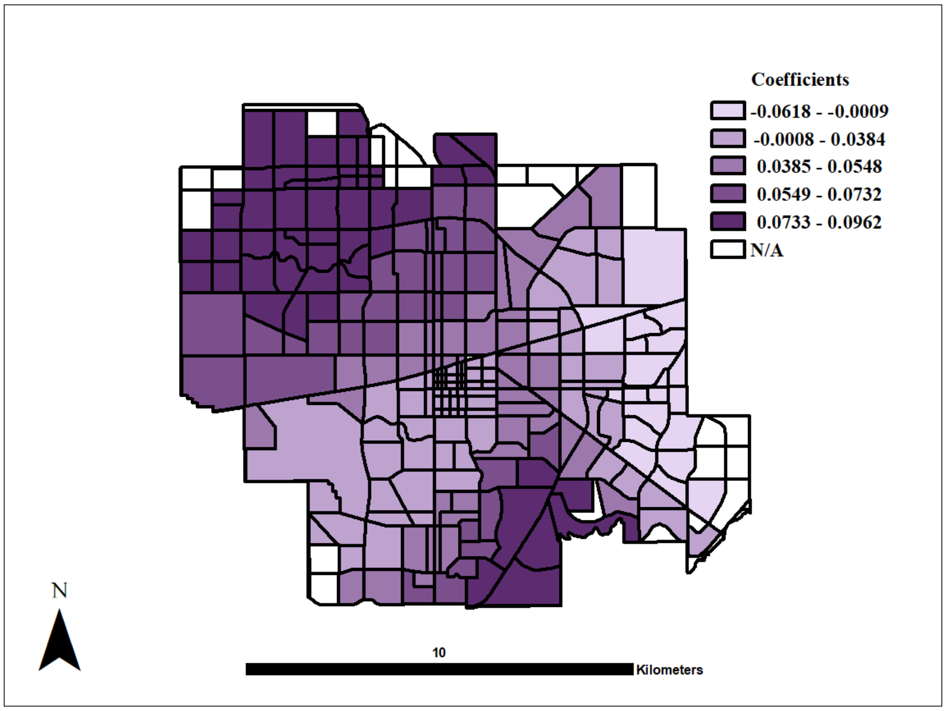
F. WEIGHTED AVERAGE SEGMENT LENGTH



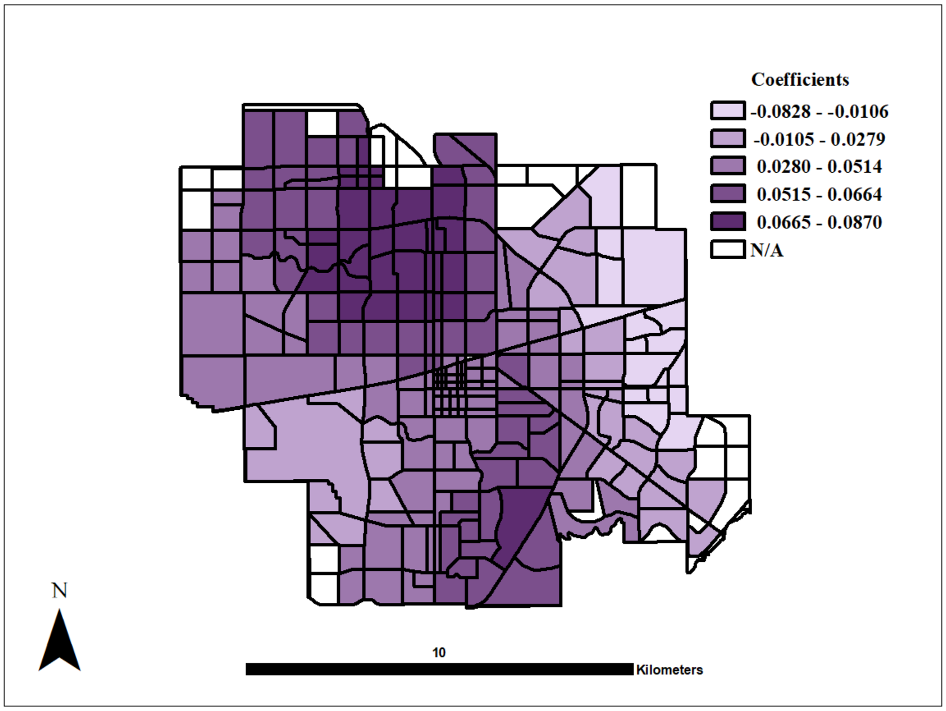
a) Fixed Gaussian Bandwidth GWPR



b) Fixed Gaussian Bandwidth GWNBR

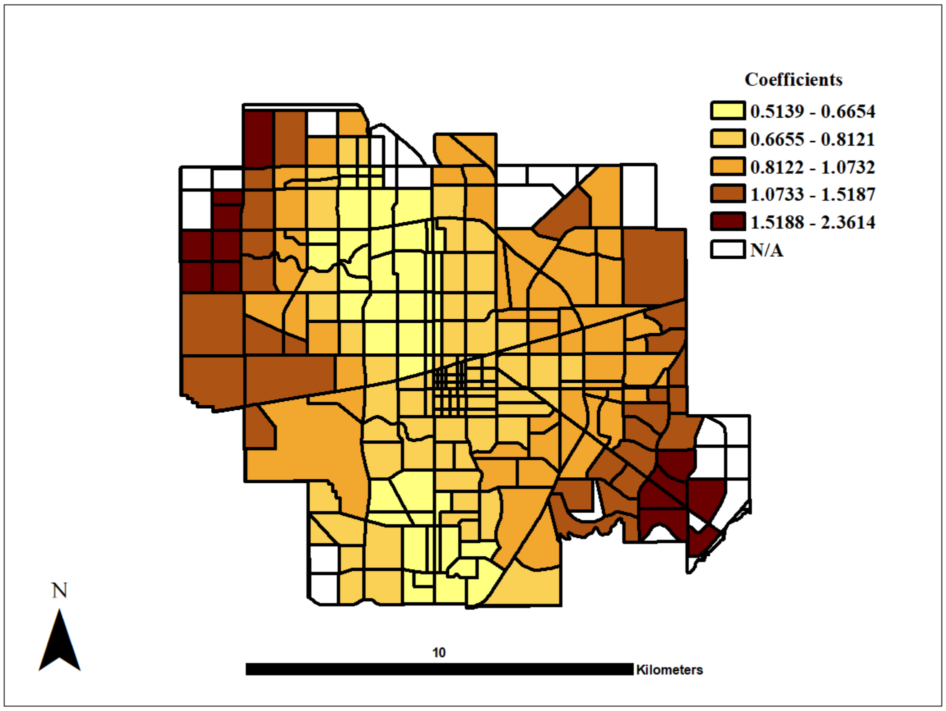


c) **Adaptive Bi-square Bandwidth GWPR**

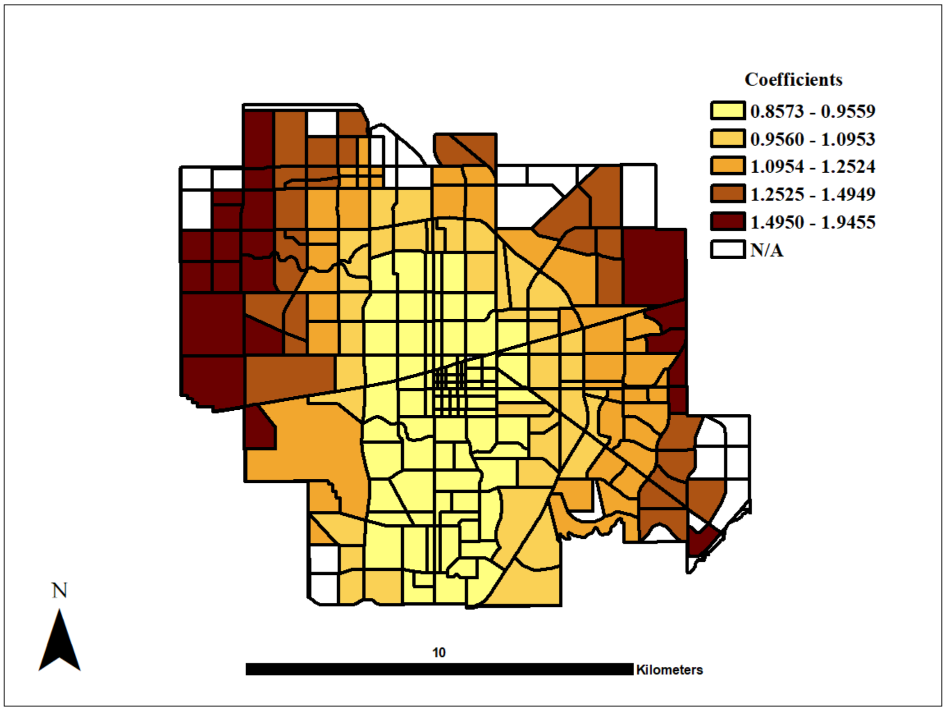


d) **Adaptive Bi-square Bandwidth GWNBR**

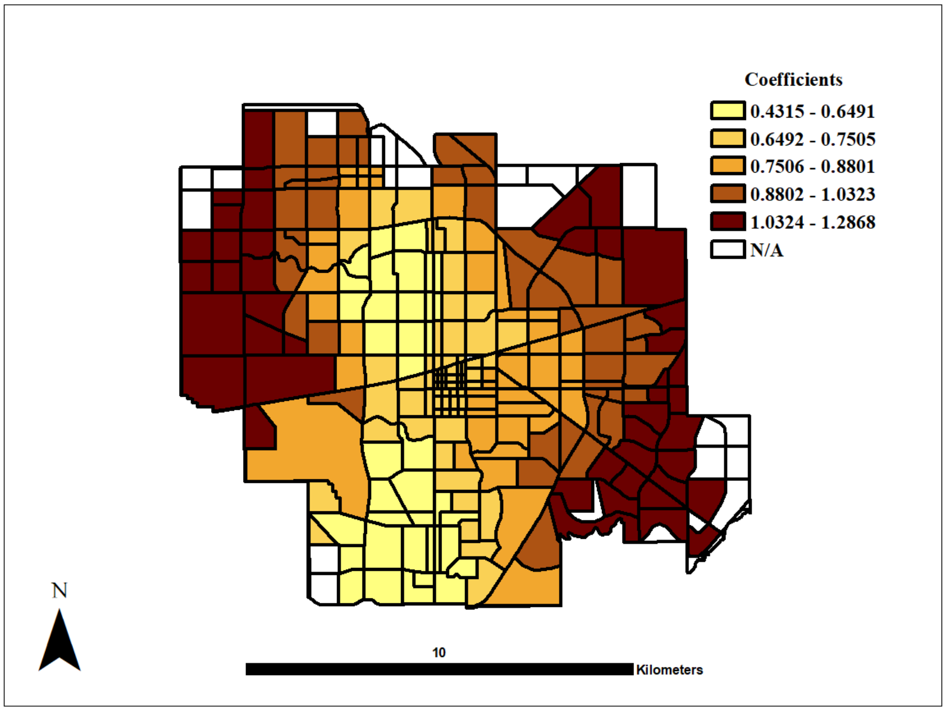
G. COMMERCIAL AREA



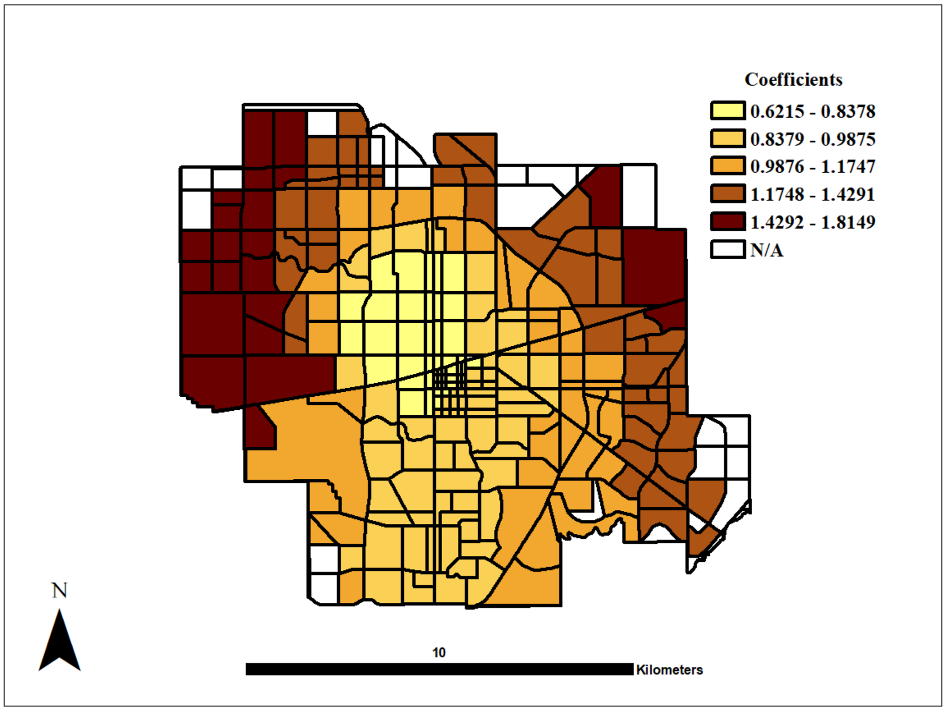
a) Fixed Gaussian Bandwidth GWPR



b) Fixed Gaussian Bandwidth GWNBR

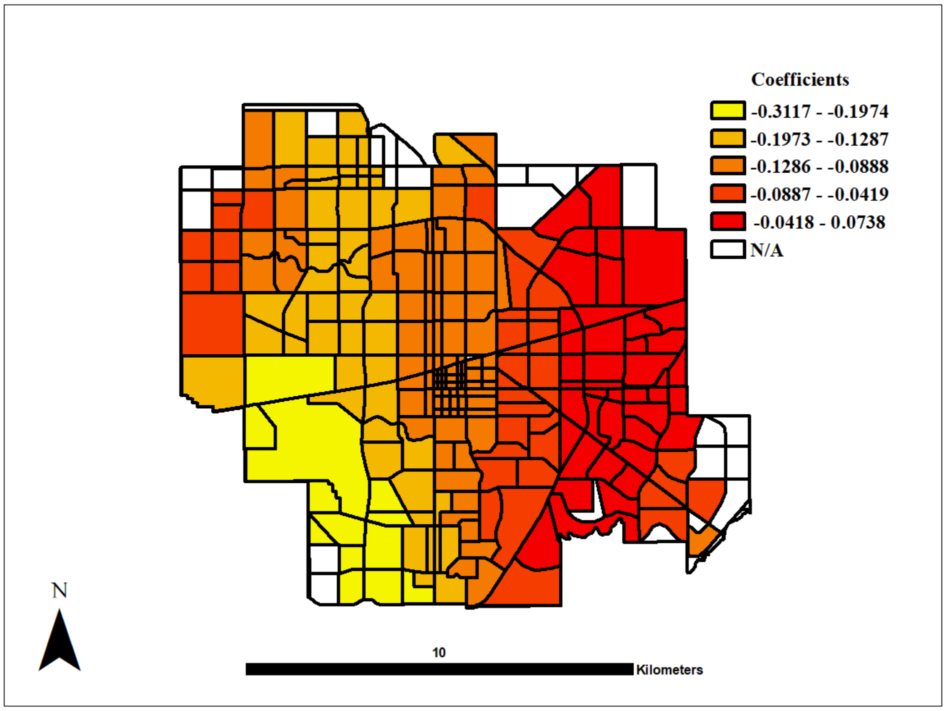


c) **Adaptive Bi-square Bandwidth GWPR**

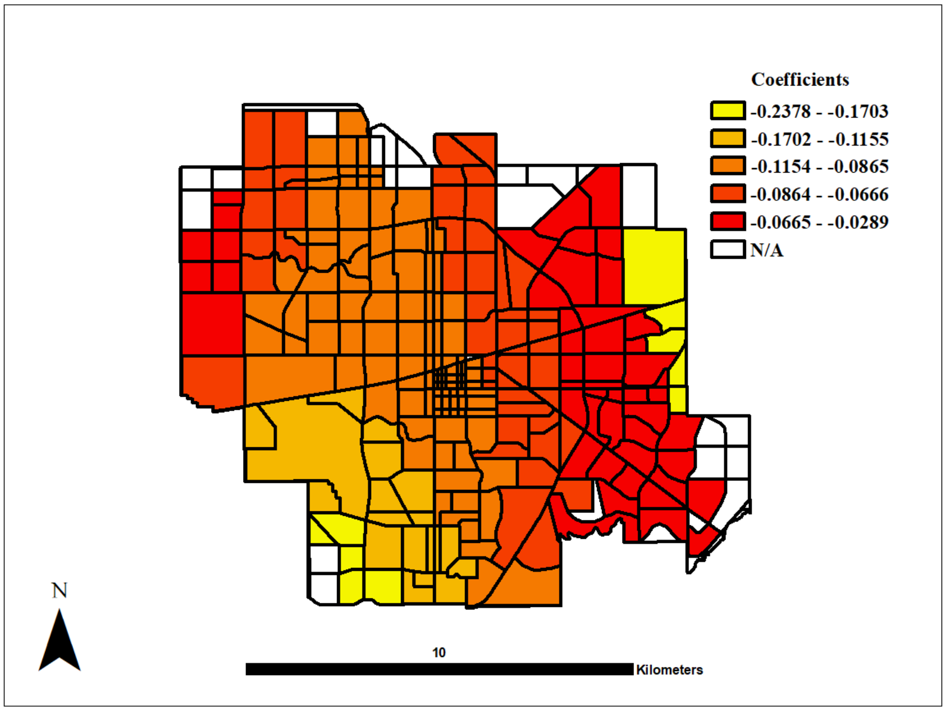


d) **Adaptive Bi-square Bandwidth GWNBR**

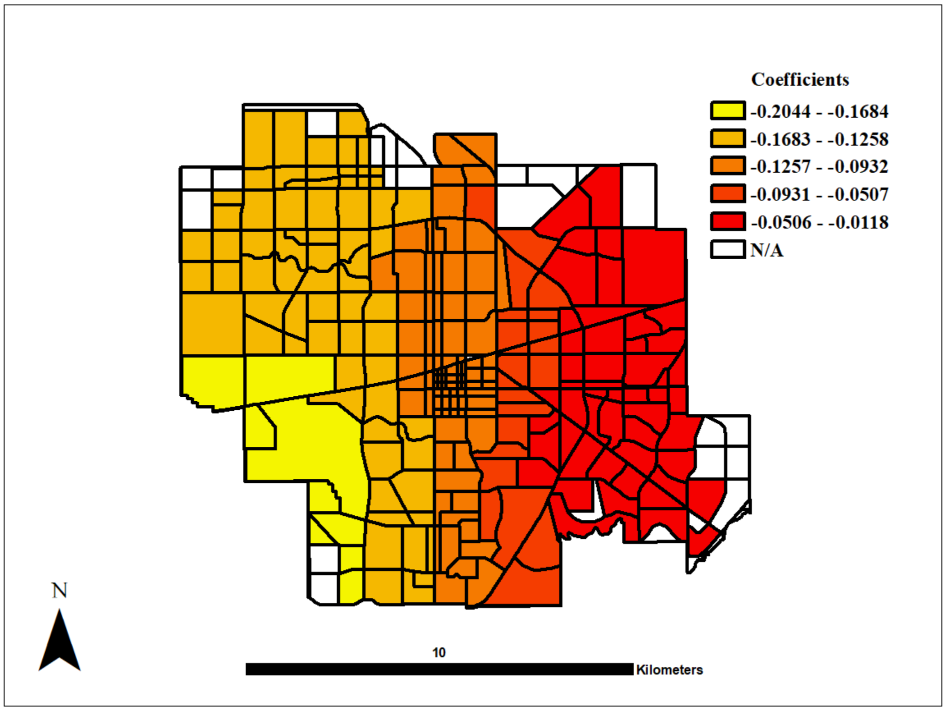
H. TOTAL SEGMENT LENGTH DENSITY



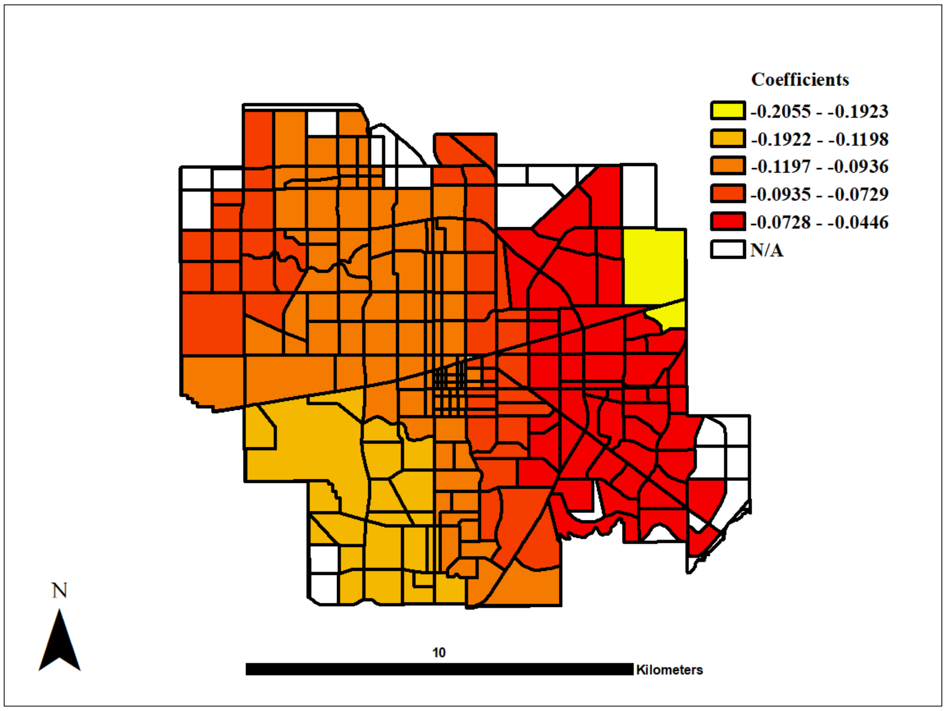
a) Fixed Gaussian Bandwidth GWPR



b) Fixed Gaussian Bandwidth GWNBR

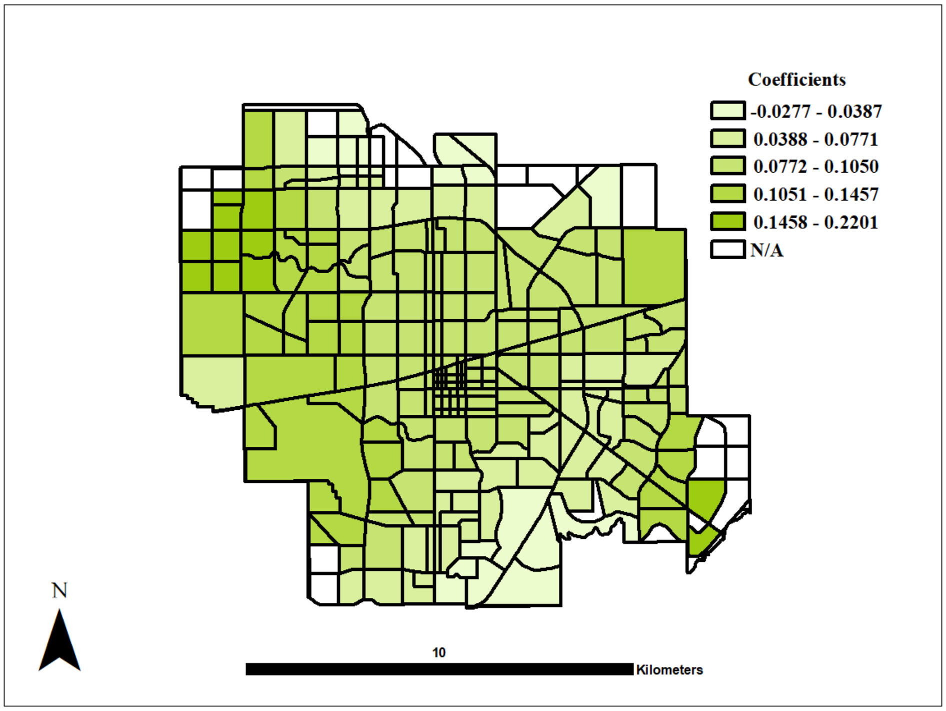


c) **Adaptive Bi-square Bandwidth GWPR**

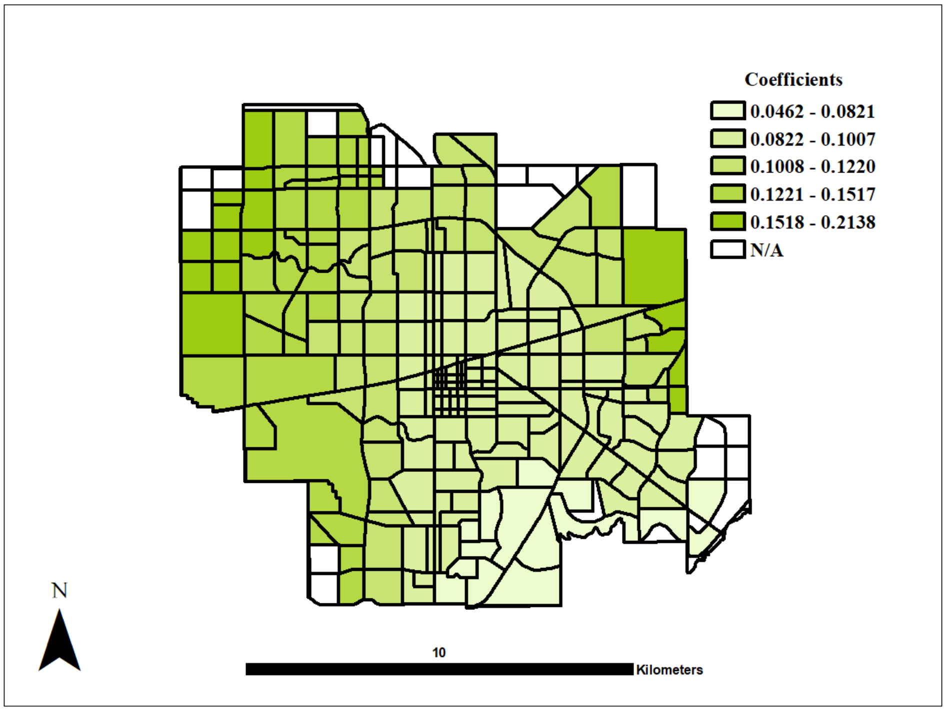


d) **Adaptive Bi- square Bandwidth GWNBR**

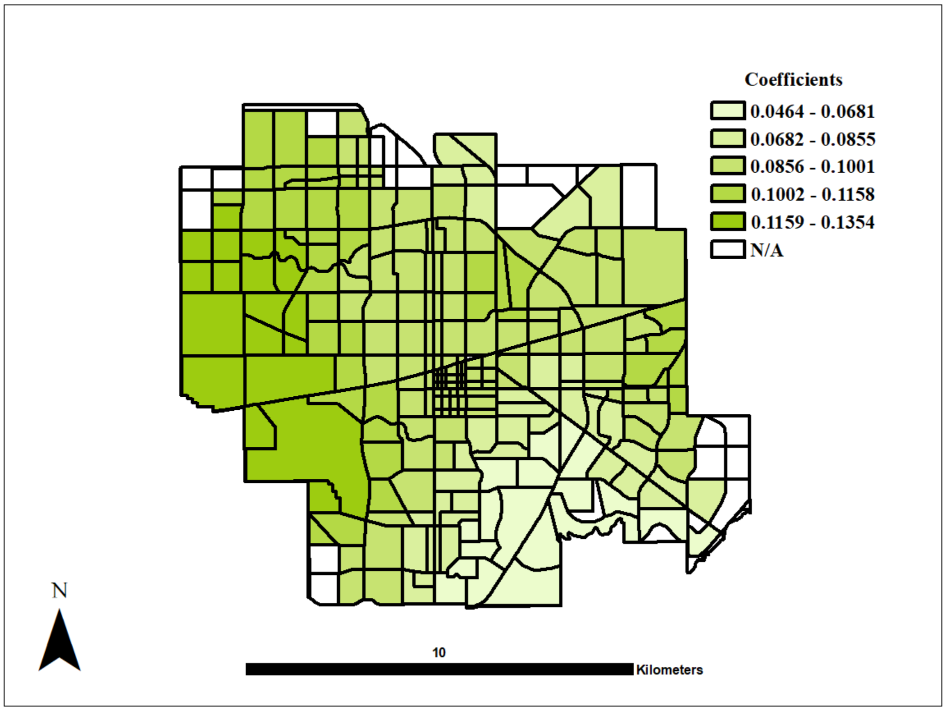
I. LOCAL ROAD DENSITY



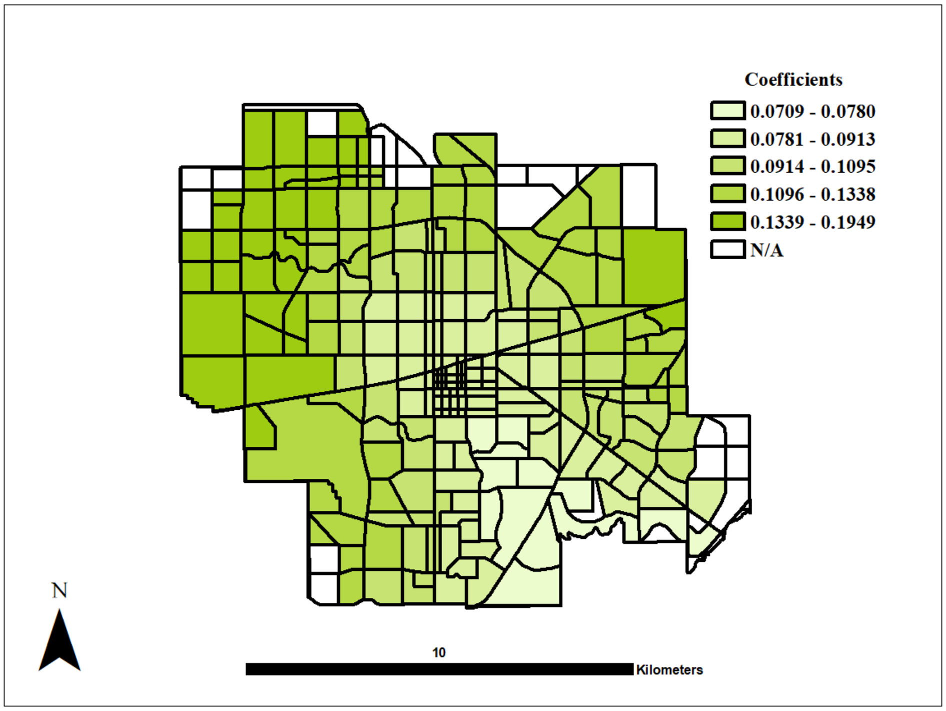
a) Fixed Gaussian Bandwidth GWPR



b) Fixed Gaussian Bandwidth GWPR



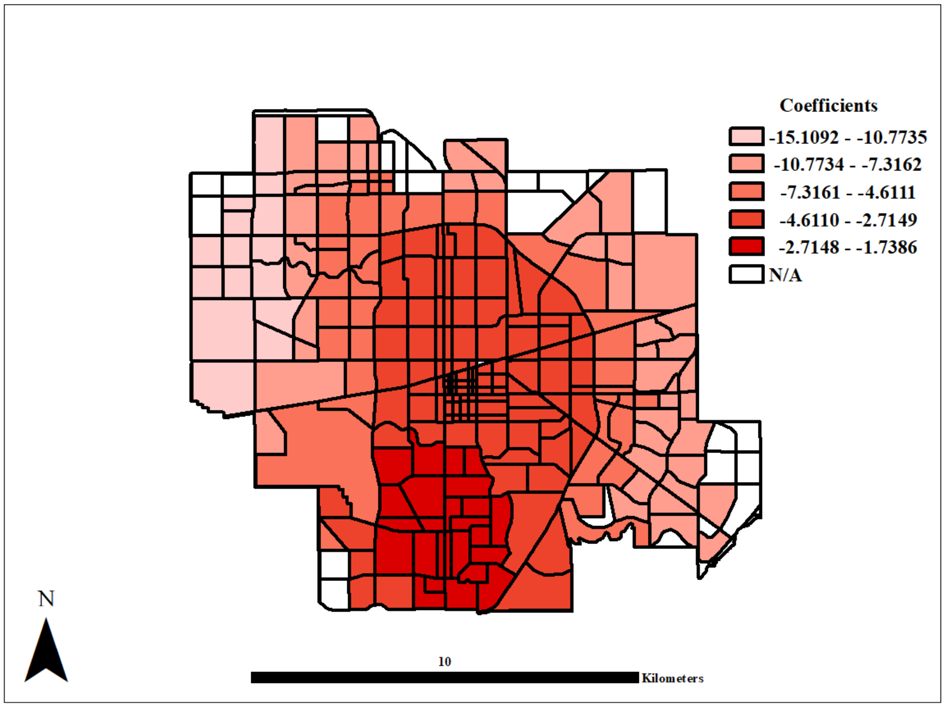
c) **Adaptive Bi-square Bandwidth GWPR**



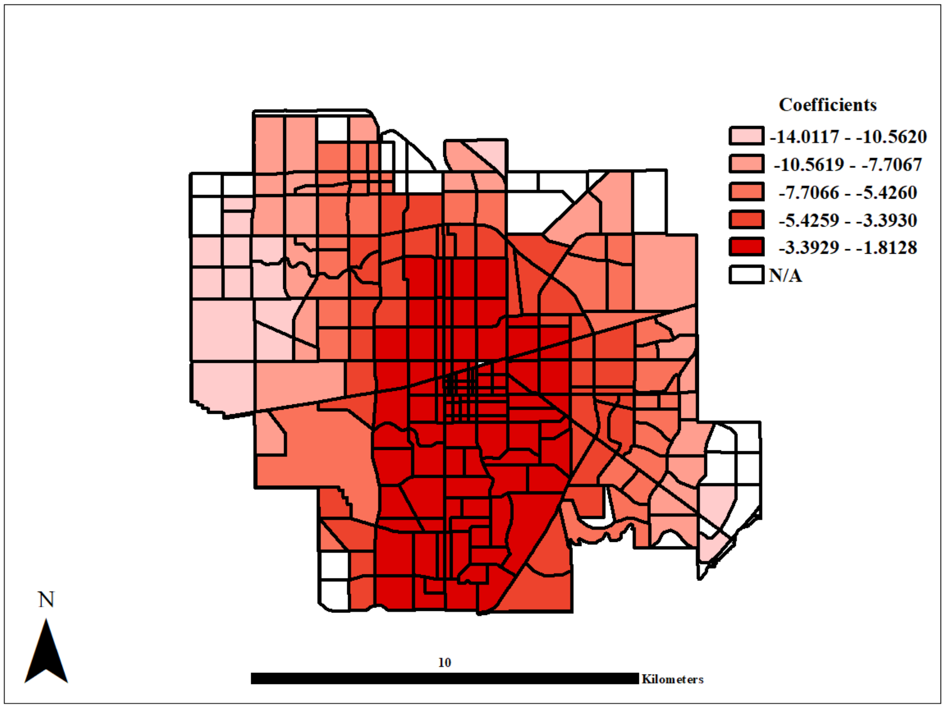
d) **Adaptive Bi-square Bandwidth GWNBR**

Appendix E: Thematic Representation of the Coefficients of GWPR and GWNBR Models for Fatal-Injury Collisions

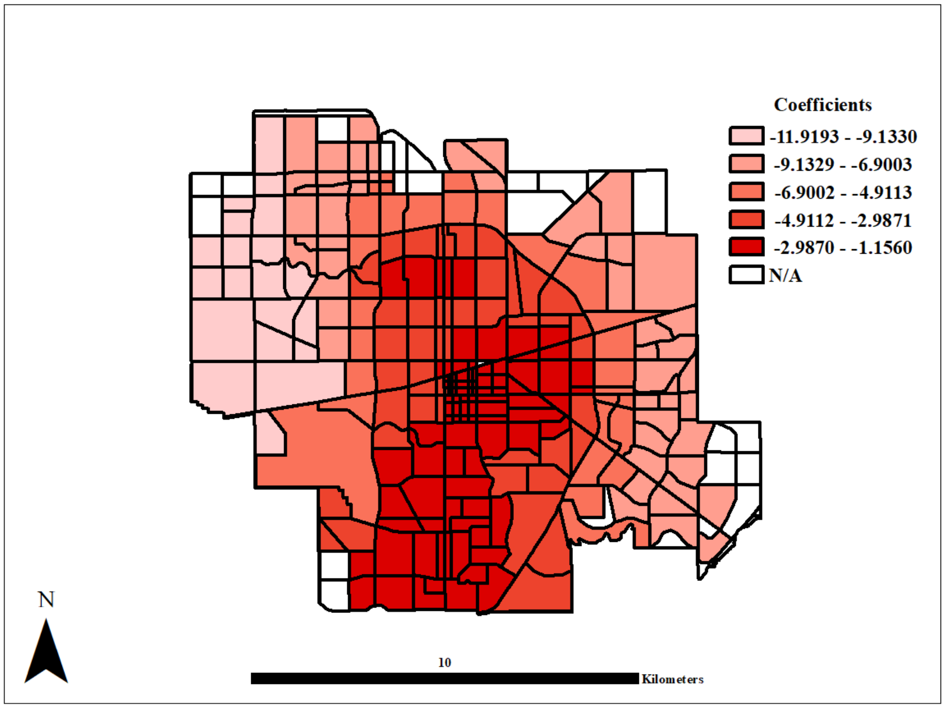
A. INTERCEPT



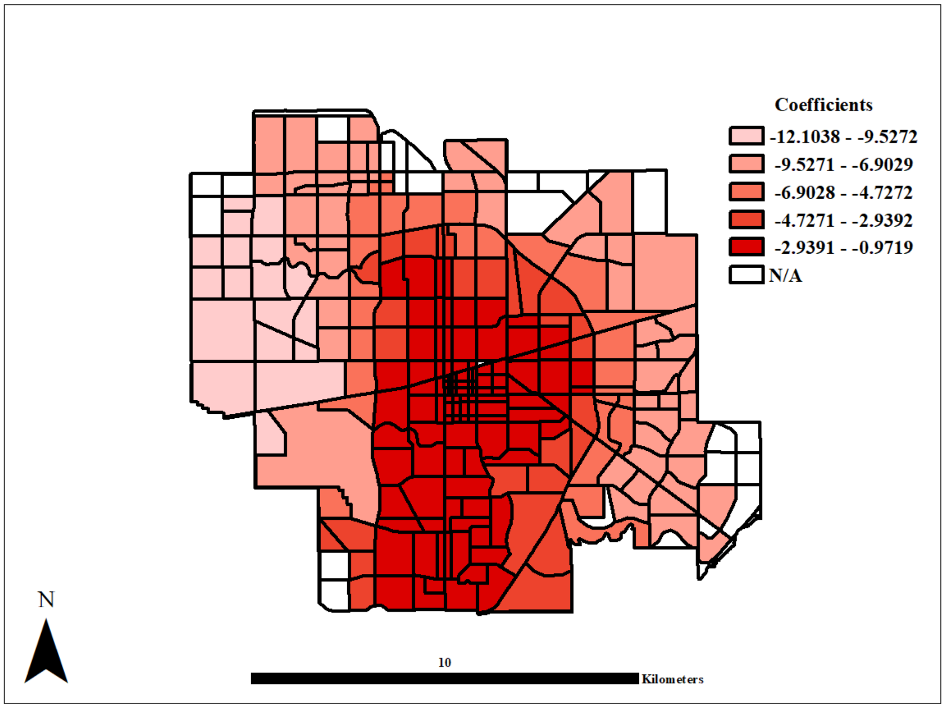
a) Fixed Gaussian Bandwidth GWPR



b) Fixed Gaussian Bandwidth GWNBR

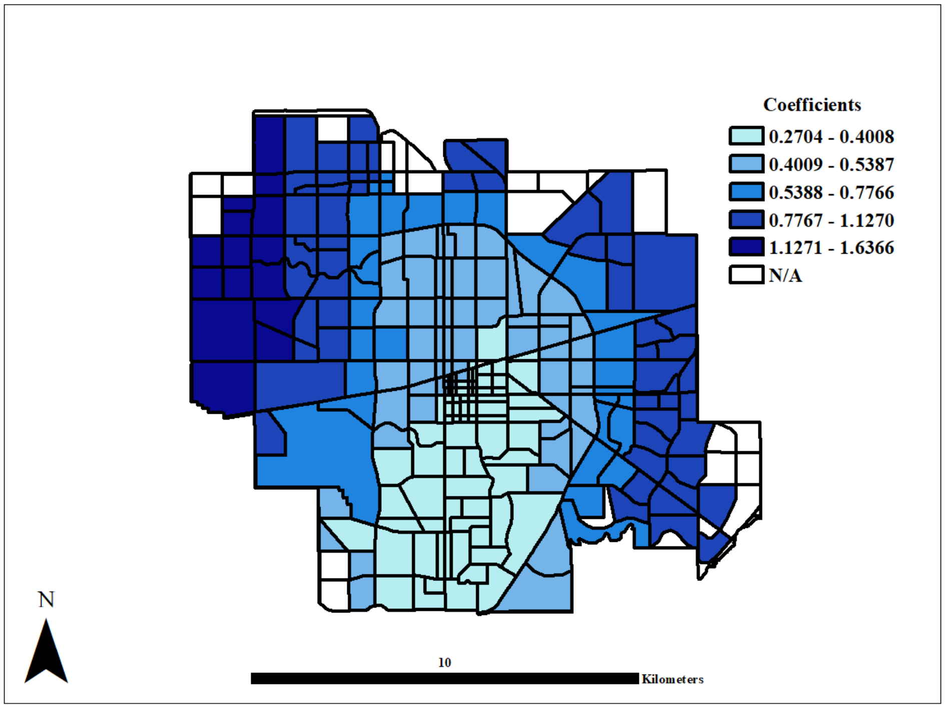


c) Adaptive Bi-Square Bandwidth GWPR

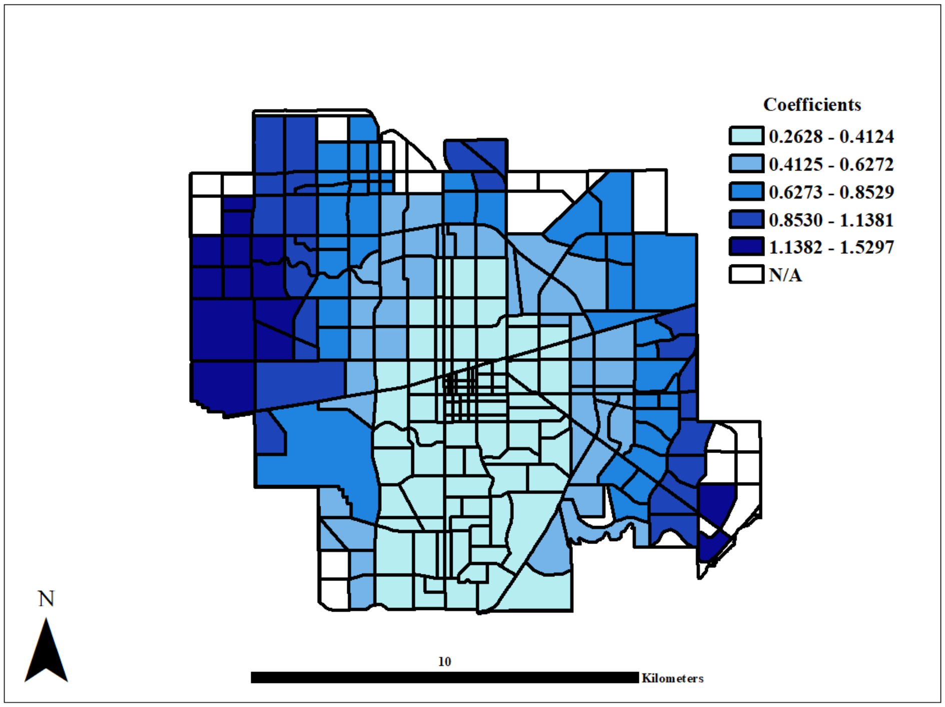


d) Adaptive Bi-square Bandwidth GWNBR

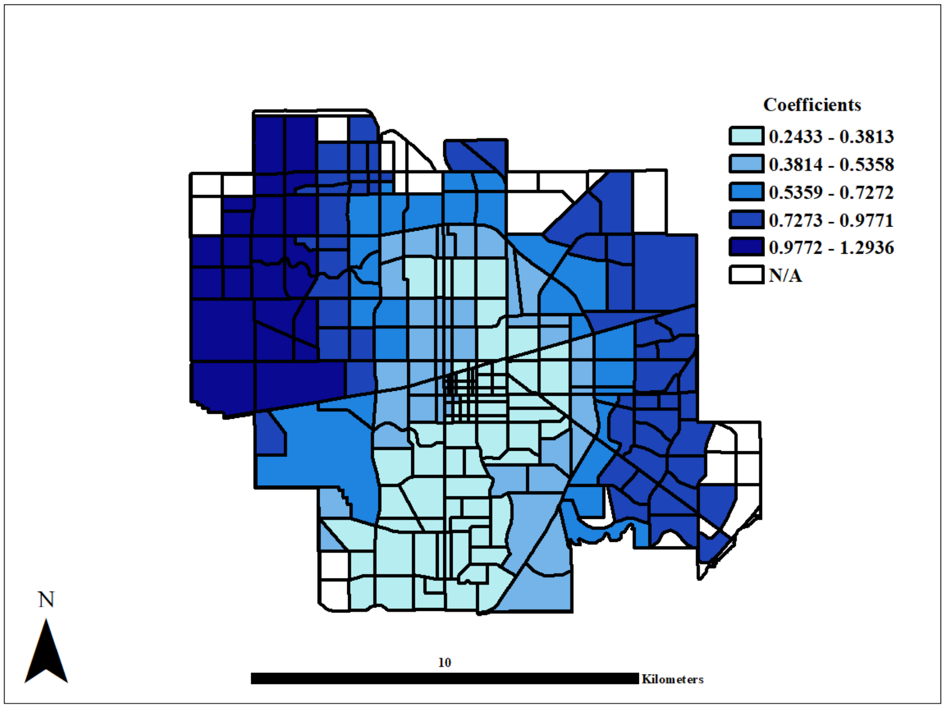
B. LOG TAZ VKMT



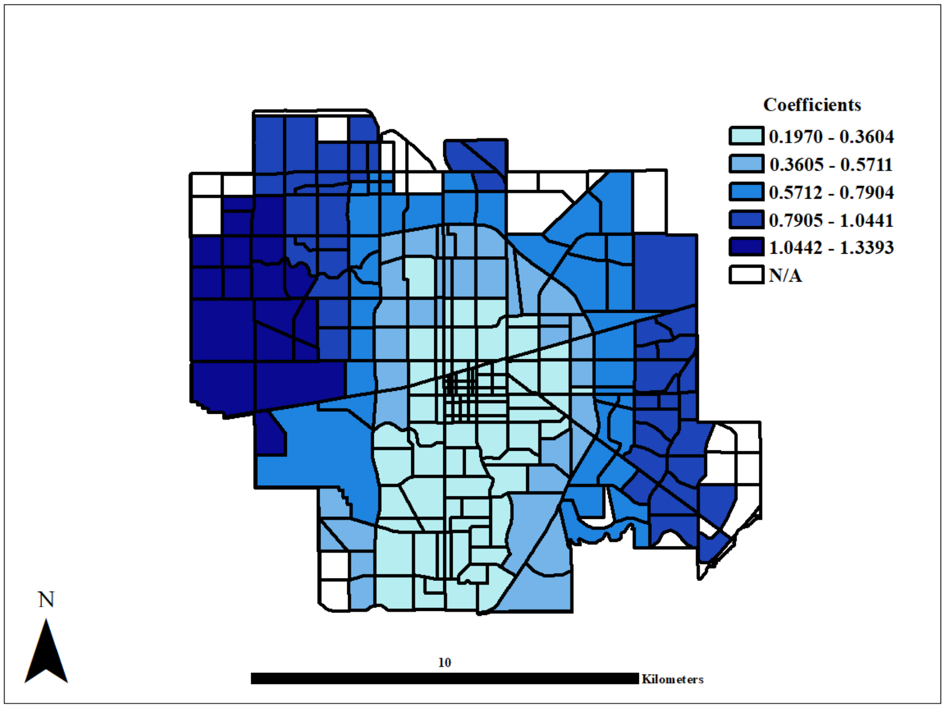
a) Fixed Gaussian Bandwidth GWPR



b) Fixed Gaussian Bandwidth GWNBR

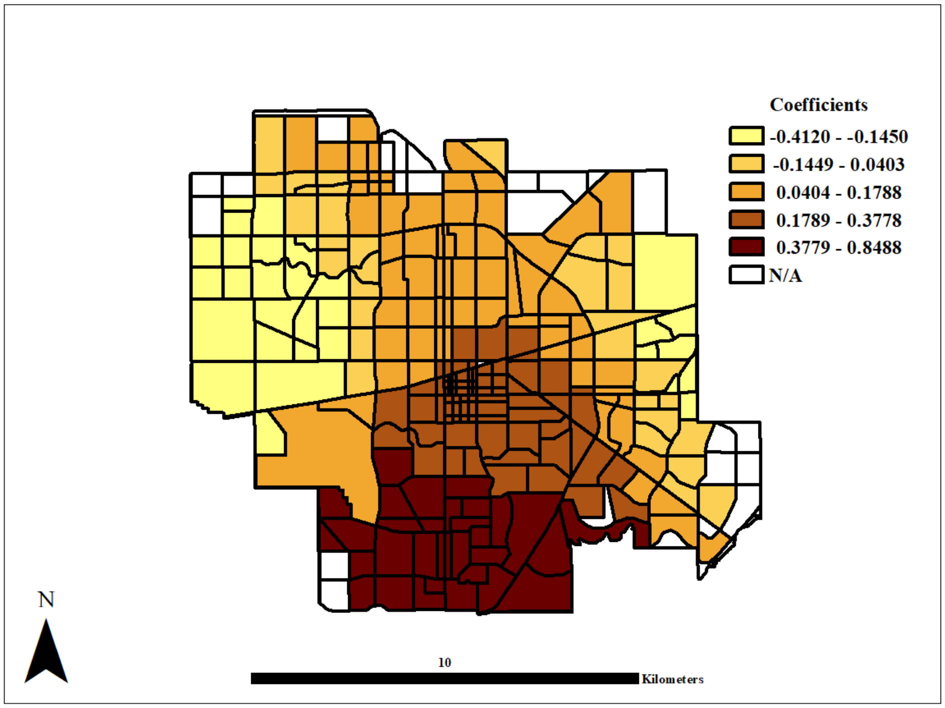


c) Adaptive Bi-square Bandwidth GWPR

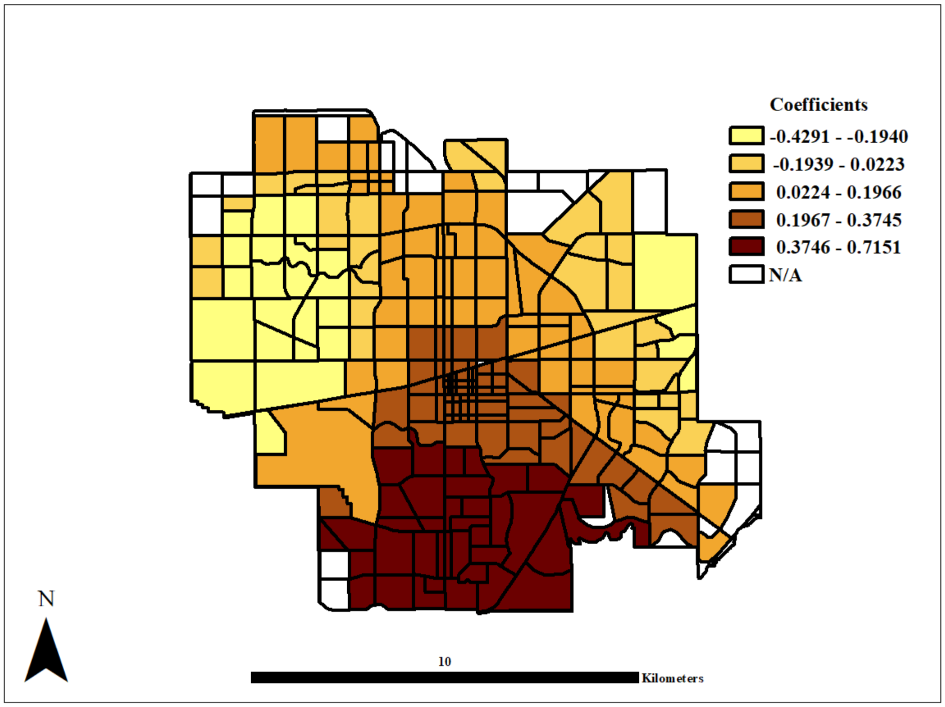


d) Adaptive Bi-square Bandwidth GWNBR

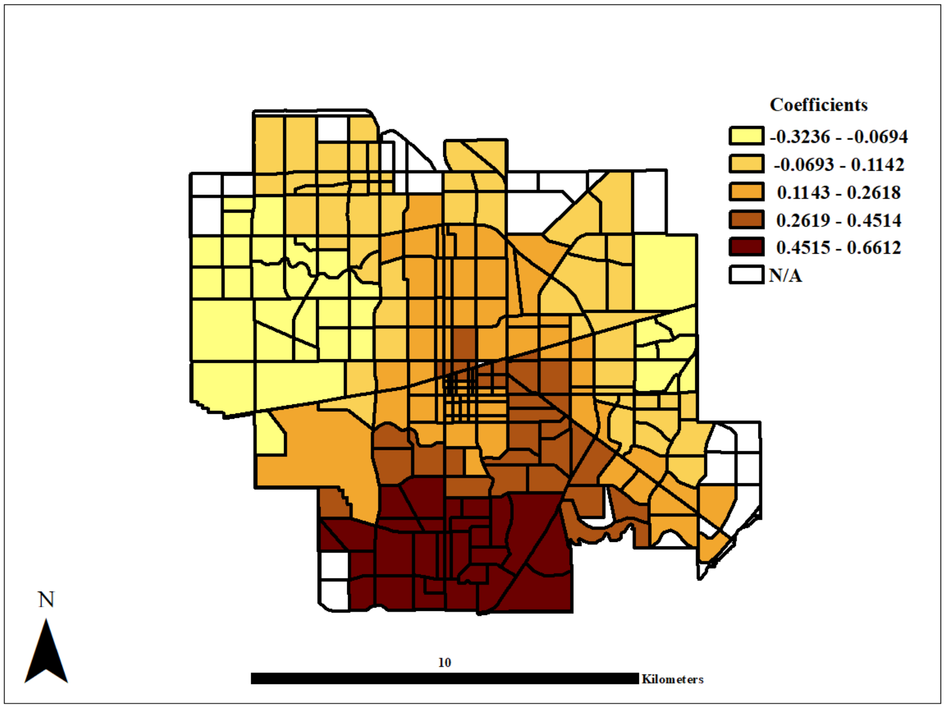
C. ARTERIAL LENGTH



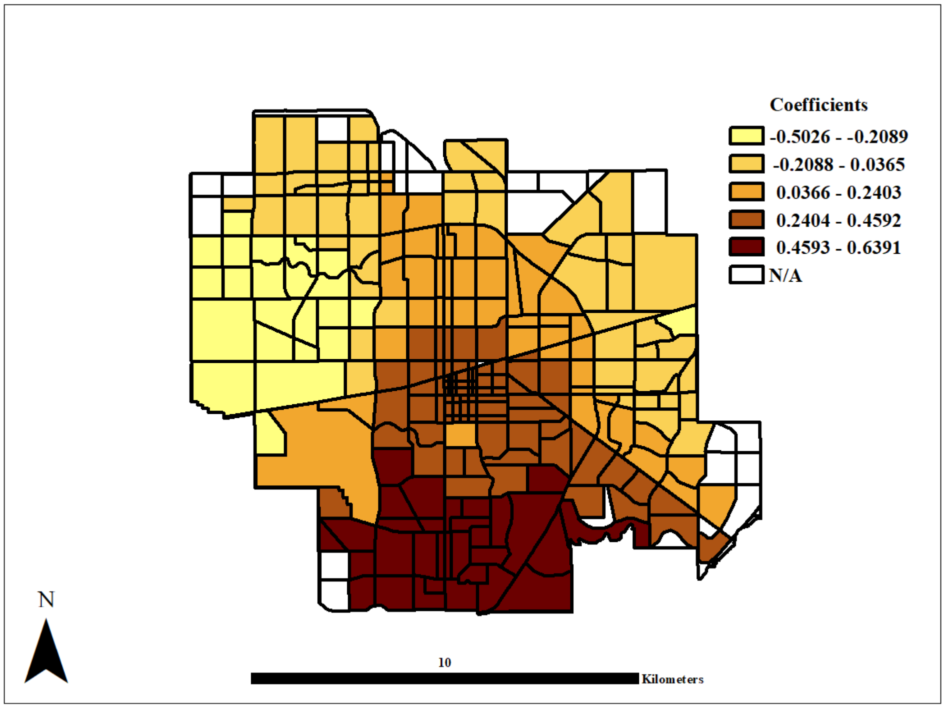
a) Fixed Gaussian Bandwidth GWPR



b) Fixed Gaussian Bandwidth GWNBR

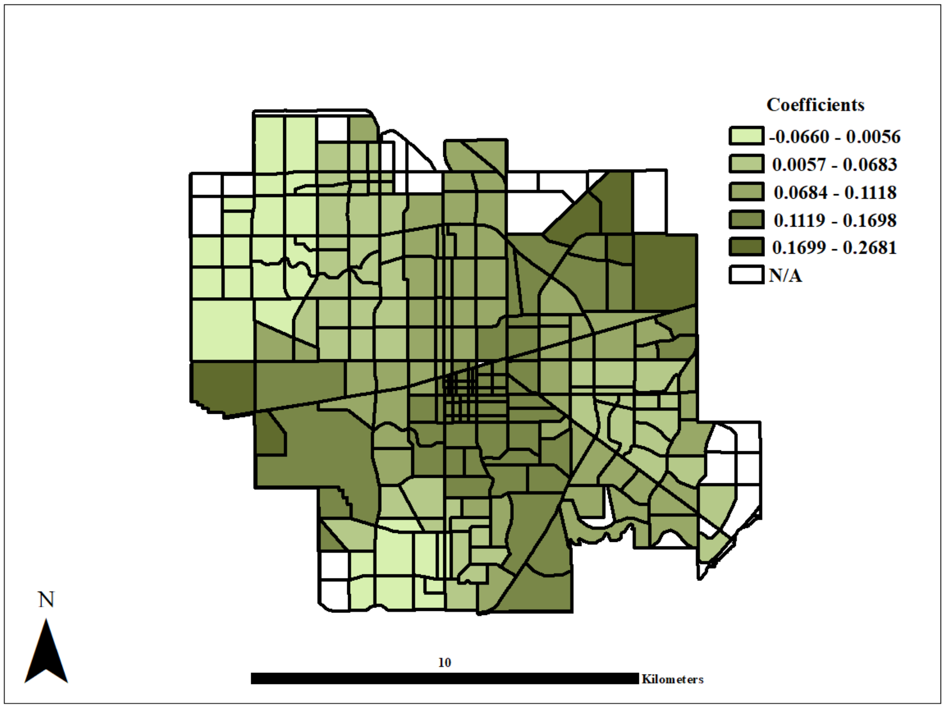


c) Adaptive Bi-square Bandwidth GWPR

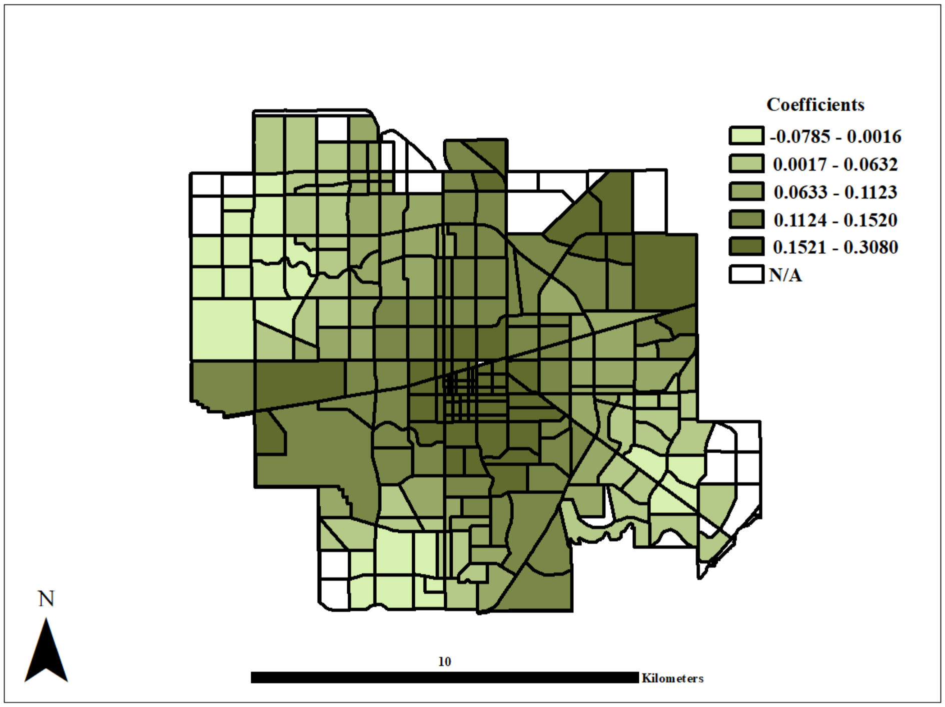


d) Adaptive Bi-square Bandwidth GWNBR

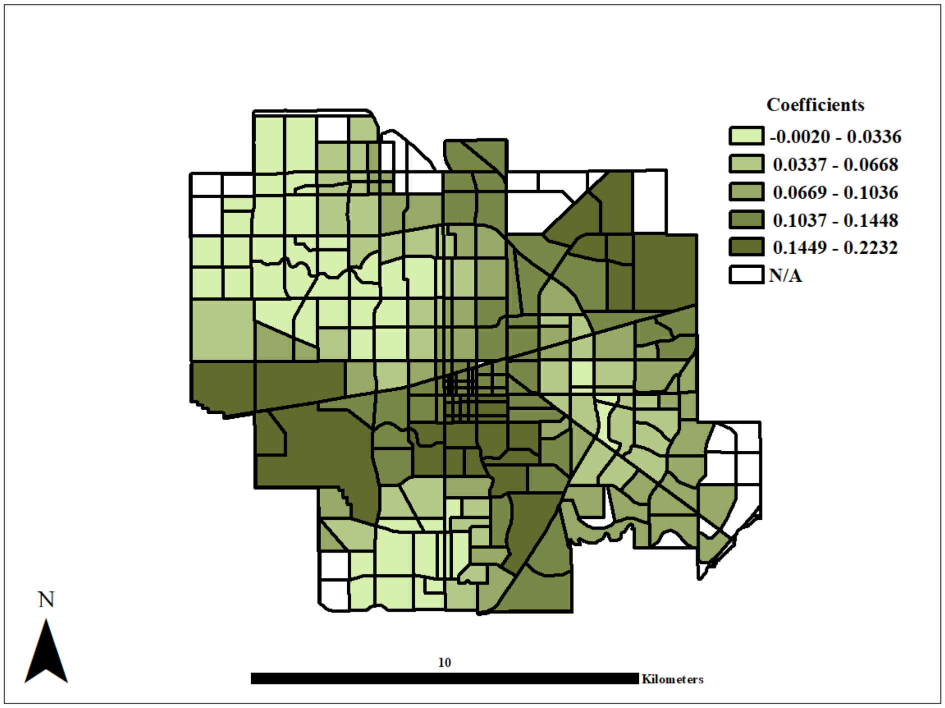
D. NUMBER OF LAND USES



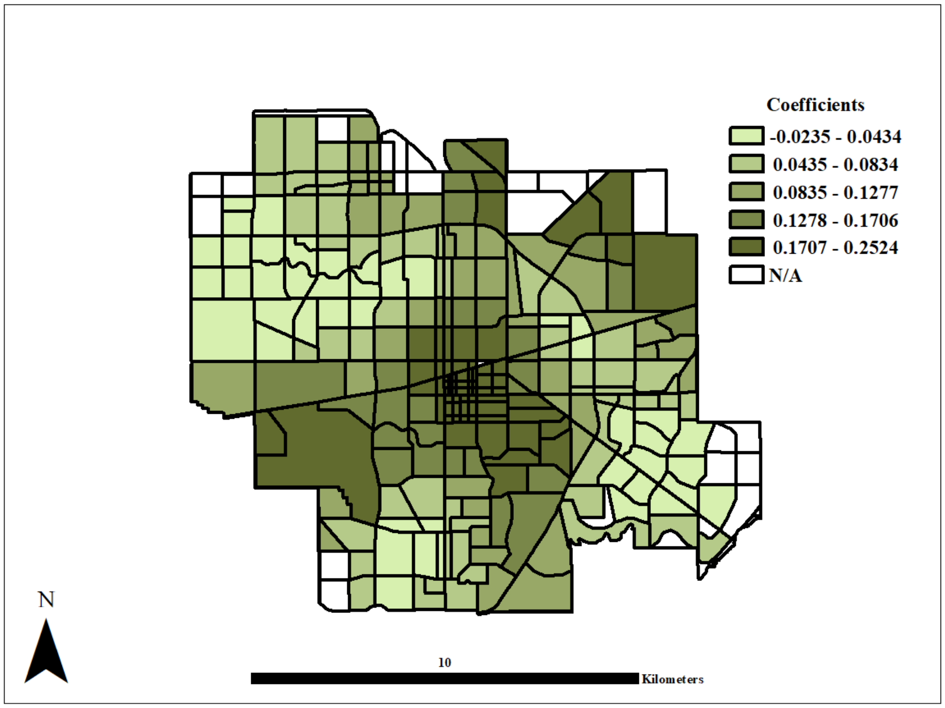
a) Fixed Gaussian Bandwidth GWPR



b) Fixed Gaussian Bandwidth GWNBR

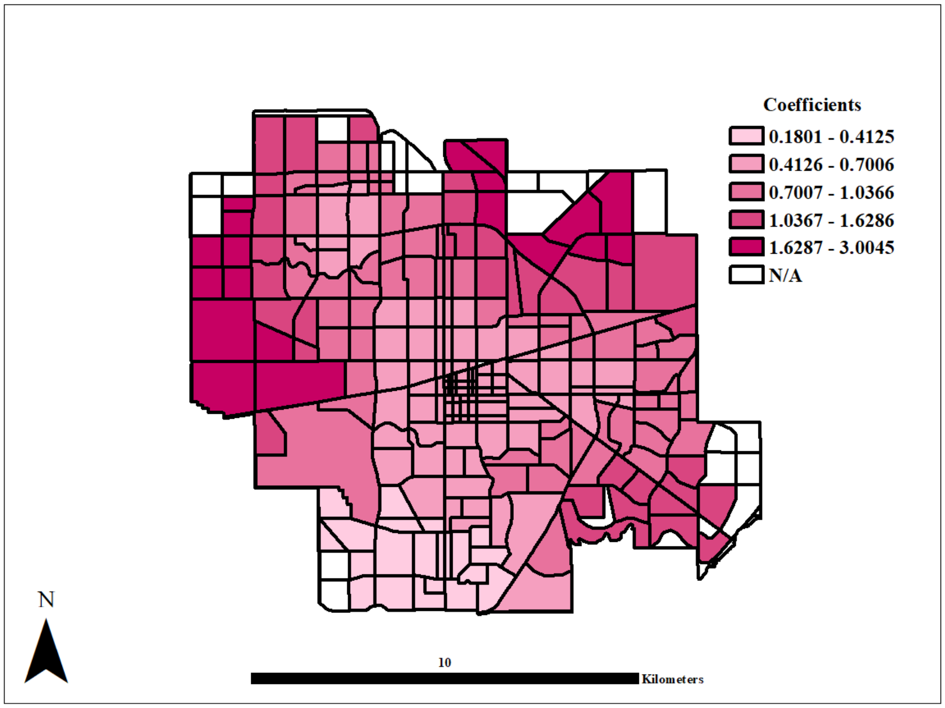


c) Adaptive Bi-square Bandwidth GWPR

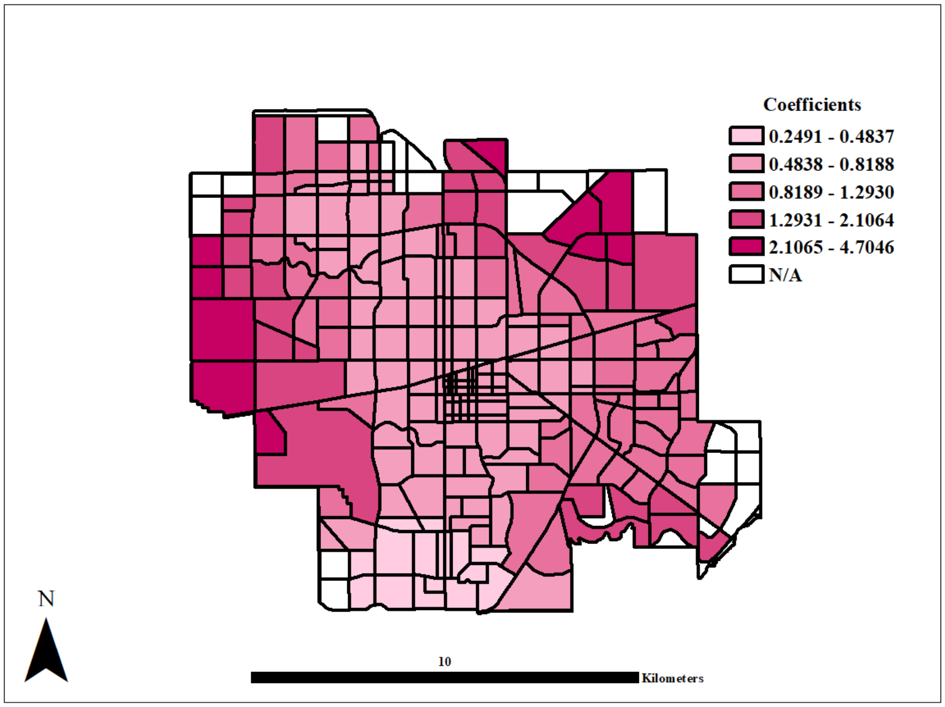


d) Adaptive Bi-square Bandwidth GWNBR

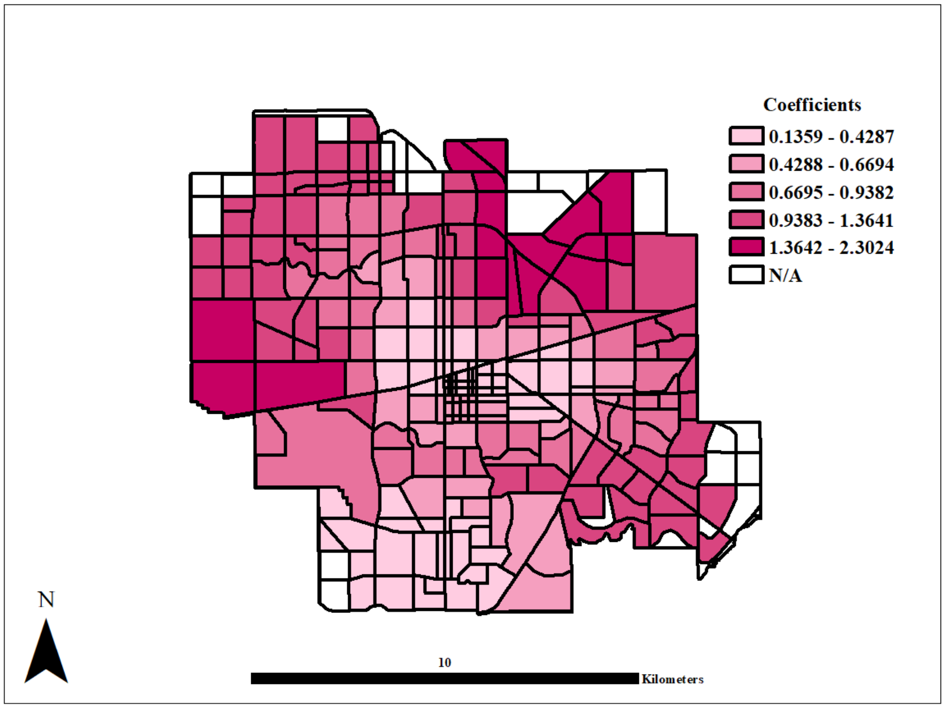
E. COMMERCIAL AREA



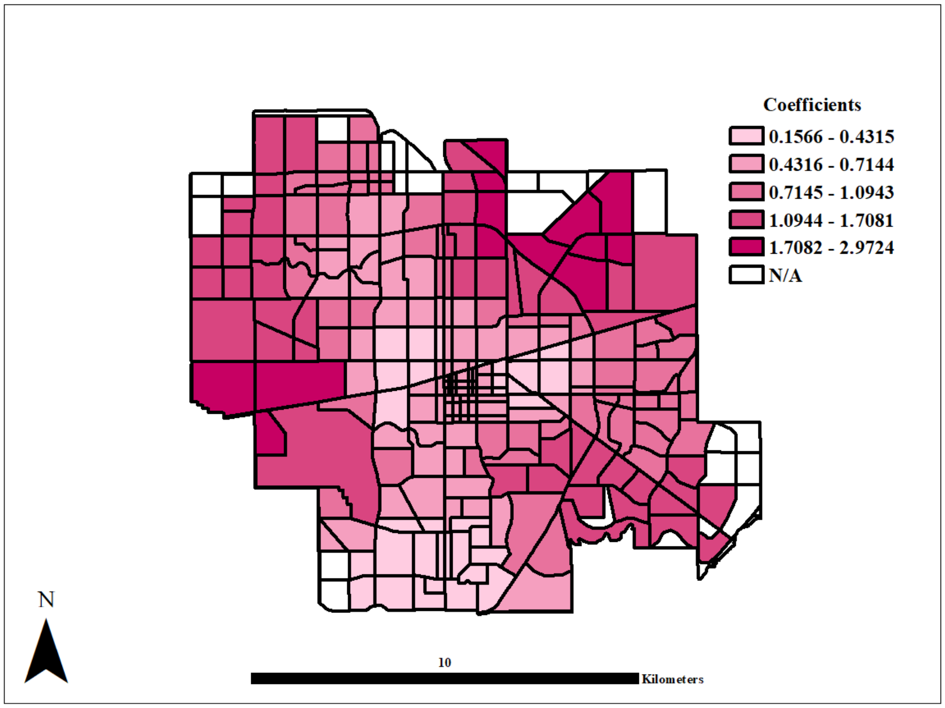
a) Fixed Gaussian Bandwidth GWPR



b) Fixed Gaussian Bandwidth GWNBR

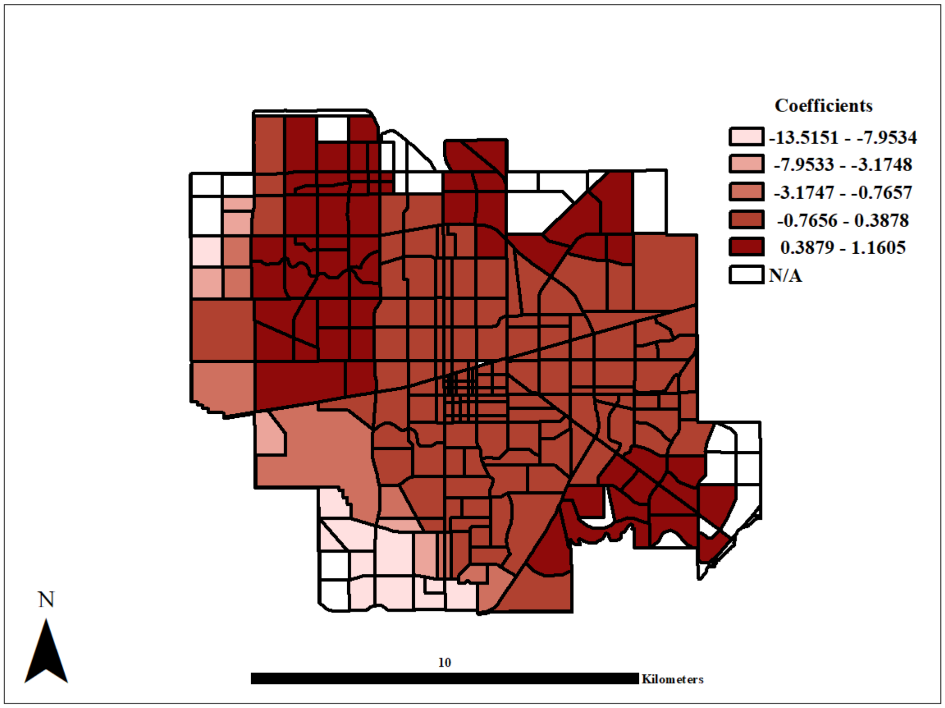


c) Adaptive Bi-square Bandwidth GWPR

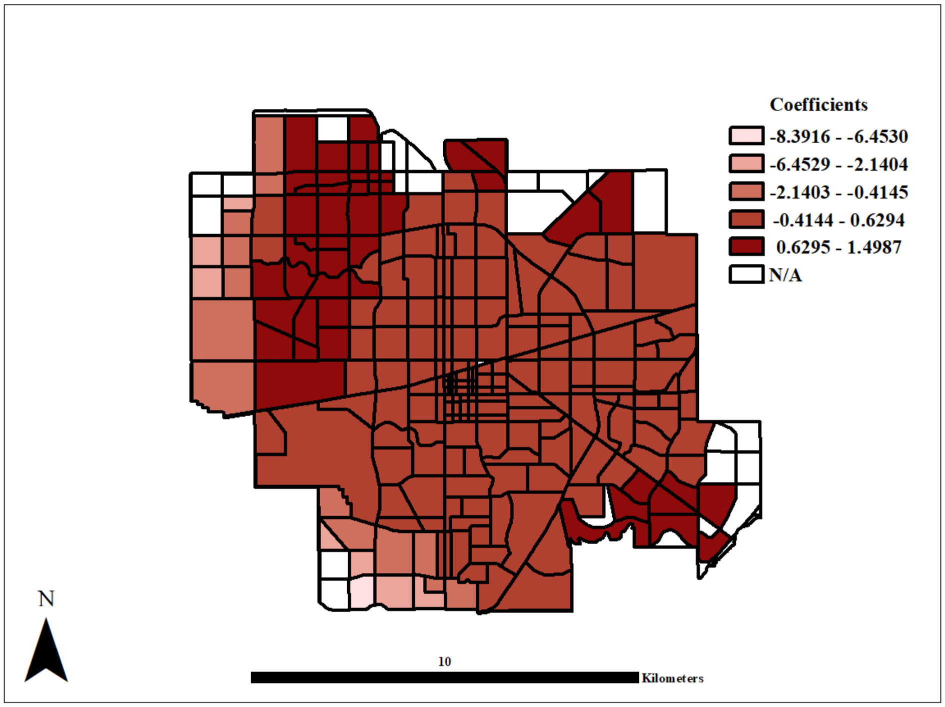


d) Adaptive Bi-square Bandwidth GWNBR

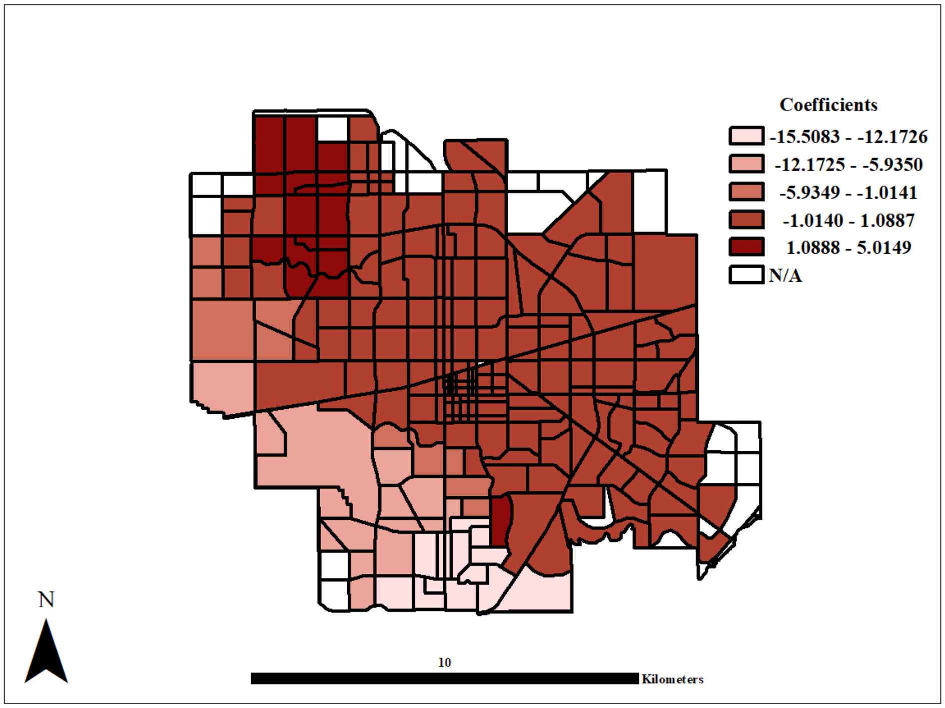
F. INDUSTRIAL AREA



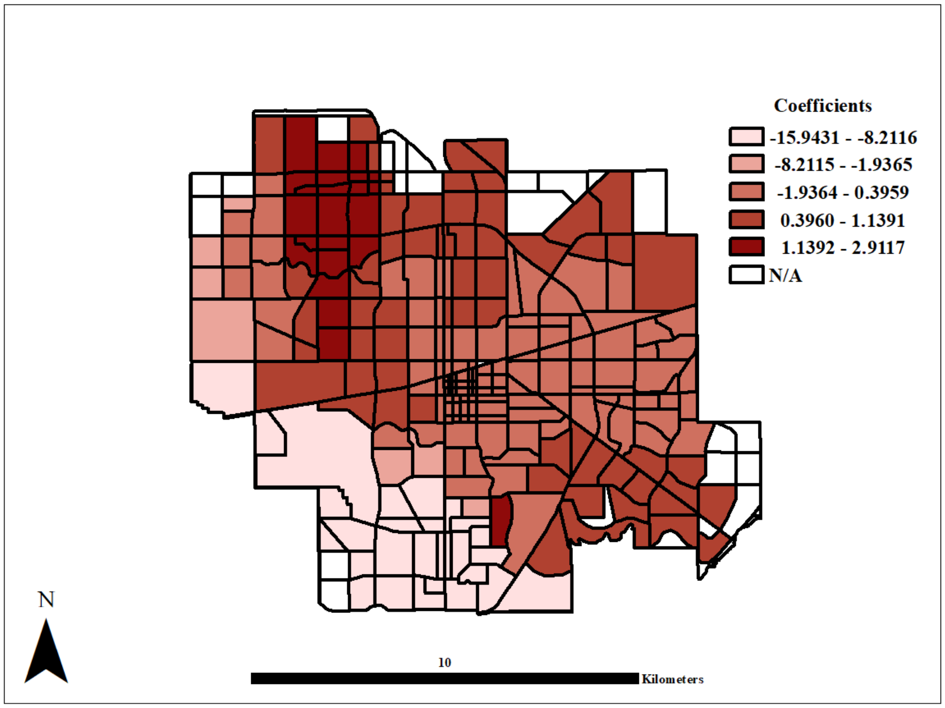
a) Fixed Gaussian Bandwidth GWPR



b) Fixed Gaussian Bandwidth GWNBR

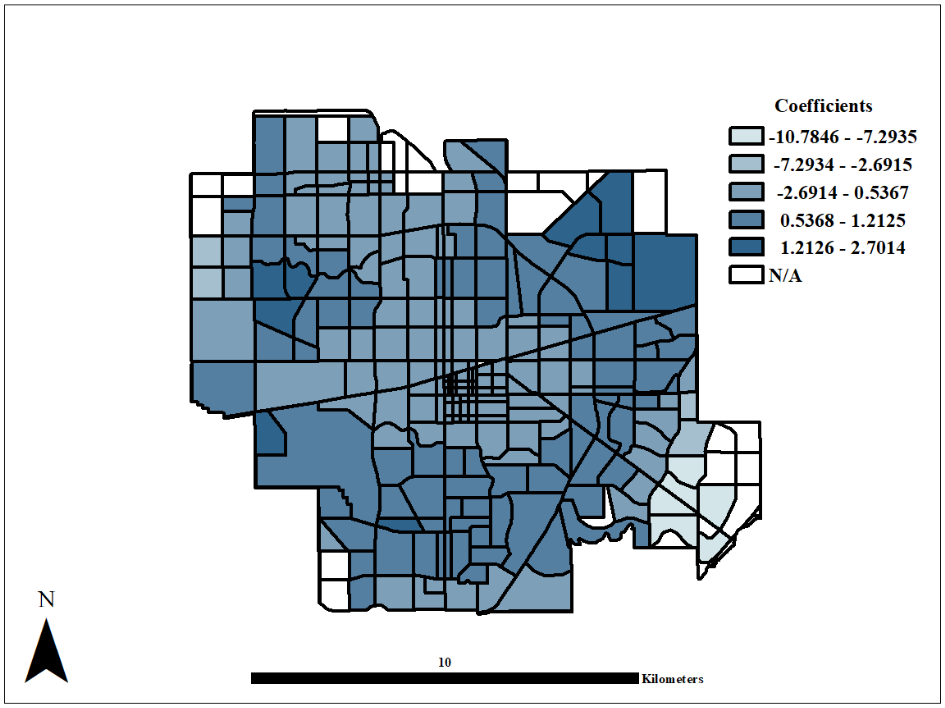


c) Adaptive Bi-square Bandwidth GWPR

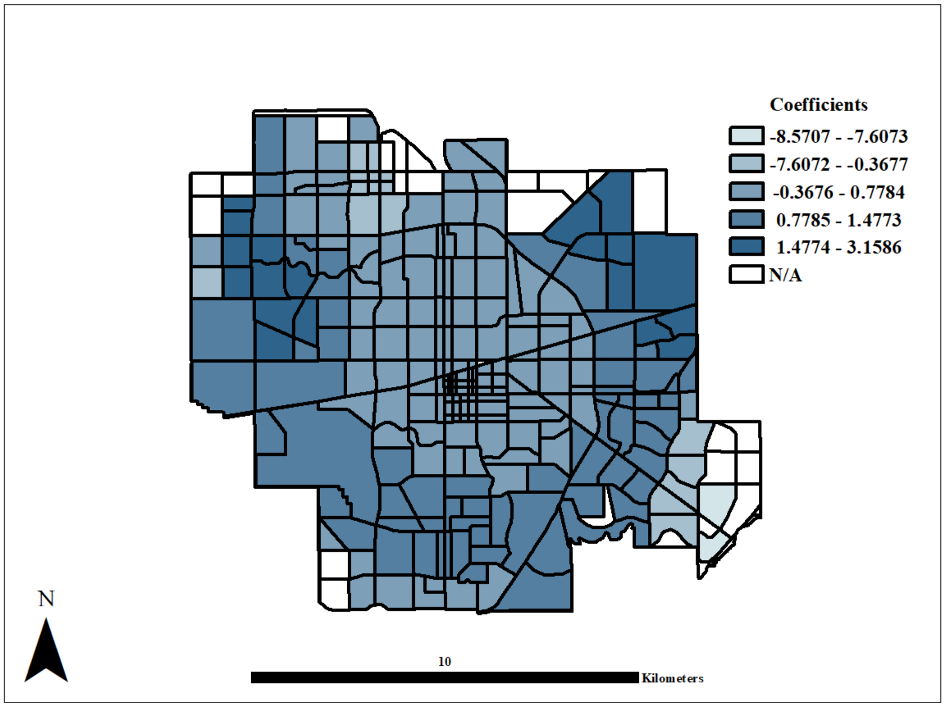


d) Adaptive Bi-square Bandwidth GWNBR

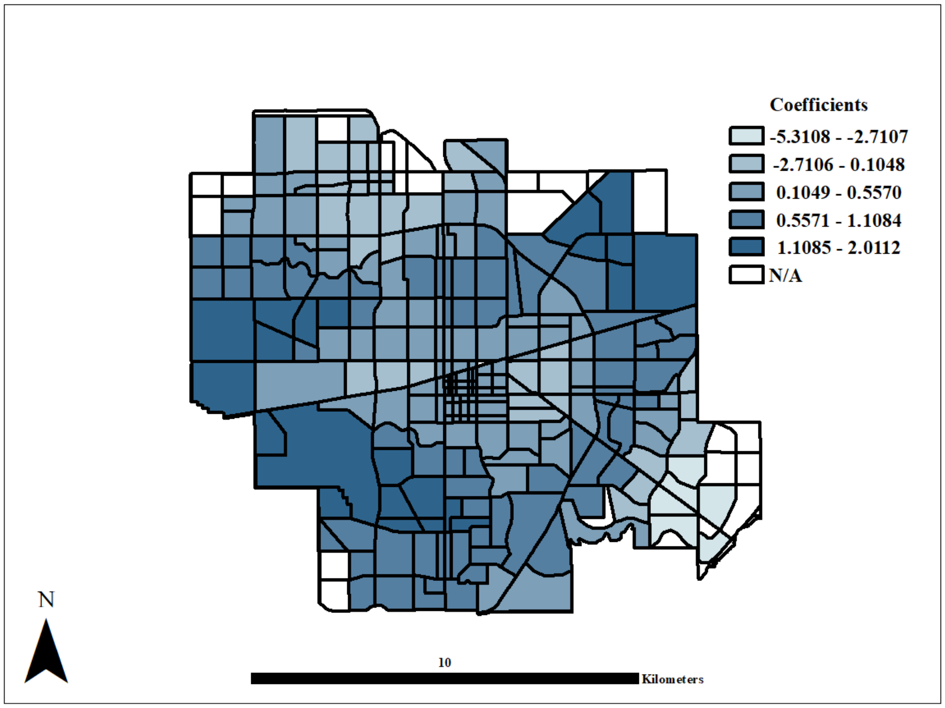
G. RESIDENTIAL MD AREA



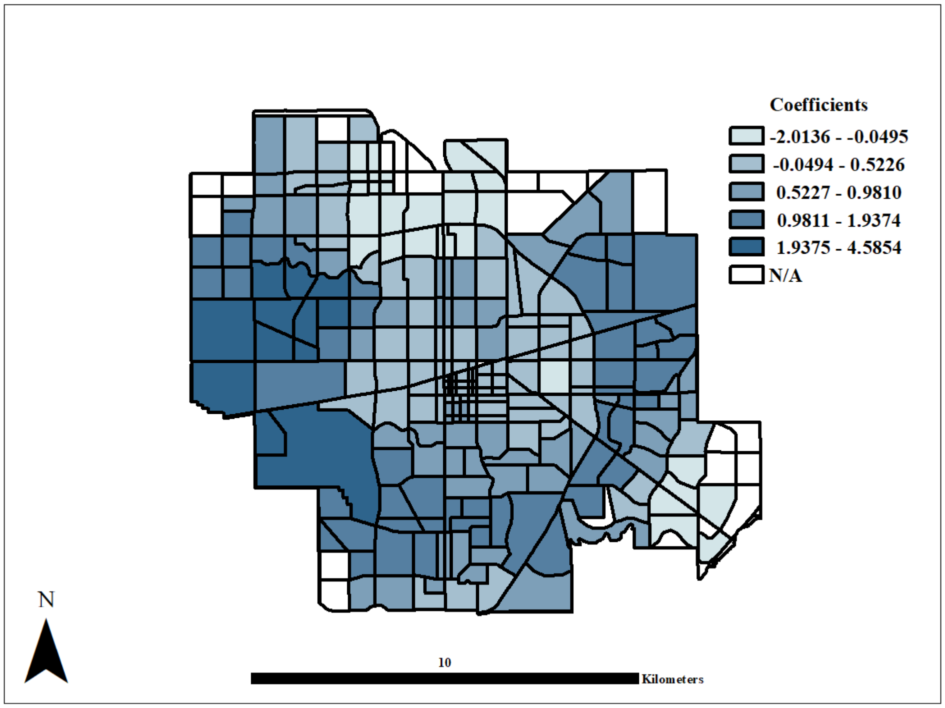
a) Fixed Gaussian Bandwidth GWPR



b) Fixed Gaussian Bandwidth GWNBR



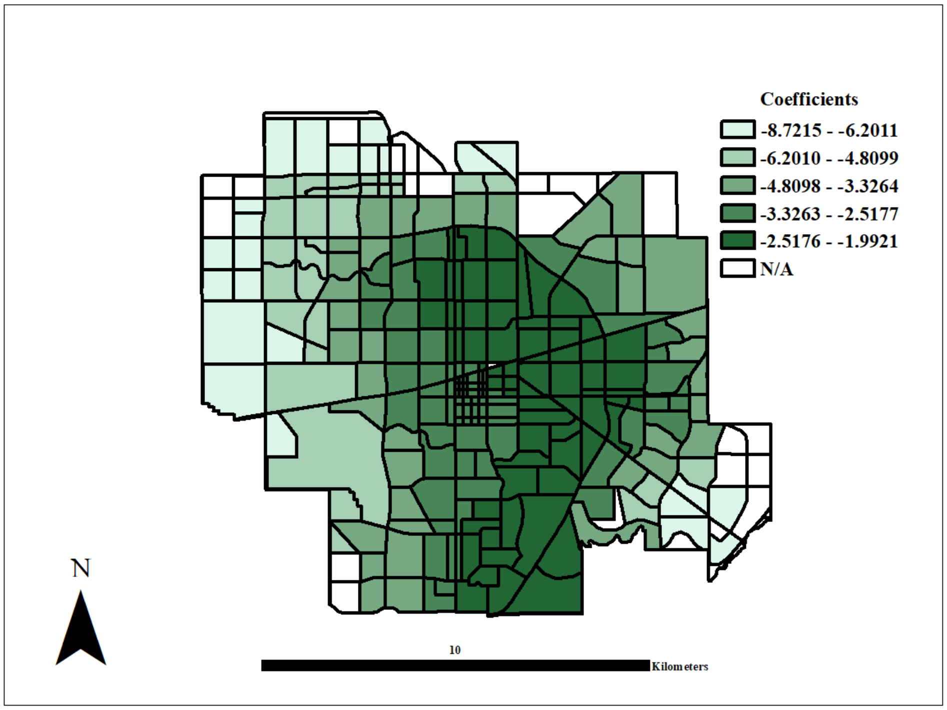
c) Adaptive Bi-square Bandwidth GWPR



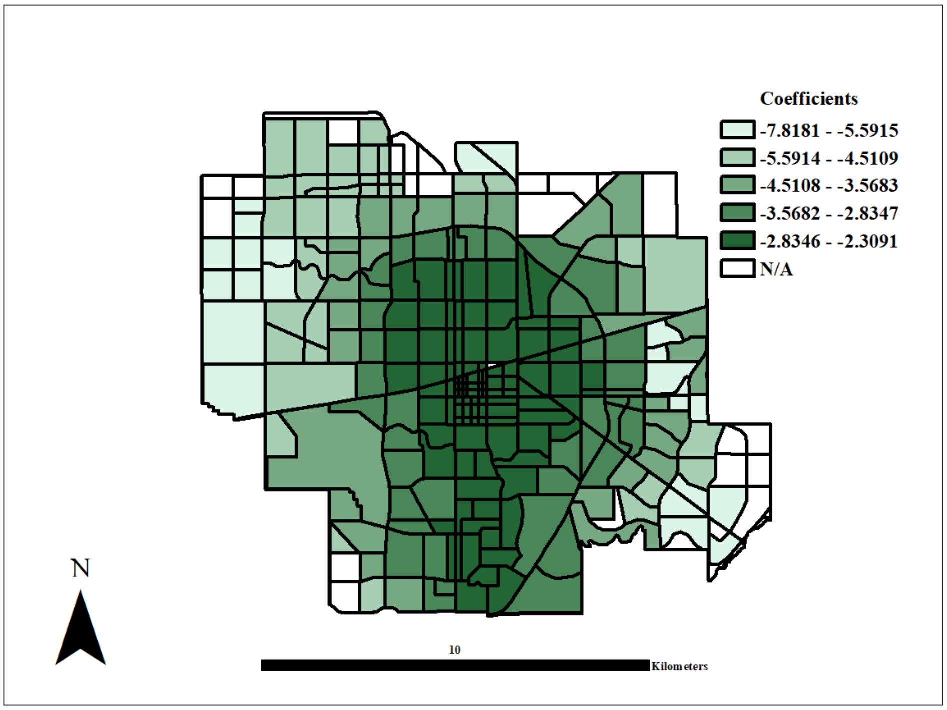
d) Adaptive Bi-square Bandwidth GWNBR

Appendix F: Thematic Representation of the Coefficients of GWPR and GWNBR Models for Property Damage Only Collisions

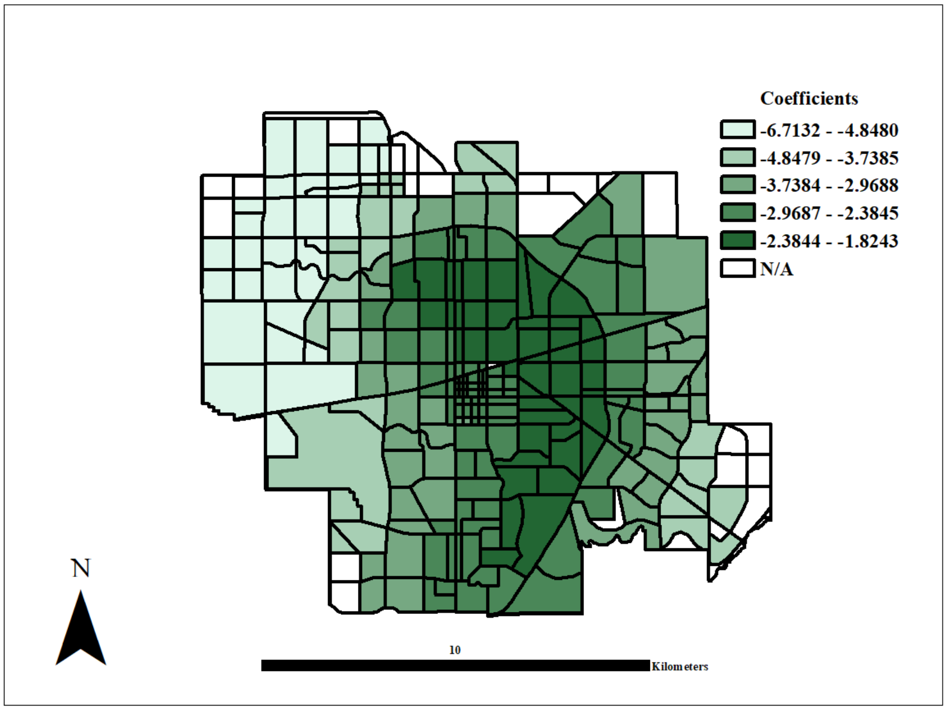
A. INTERCEPT



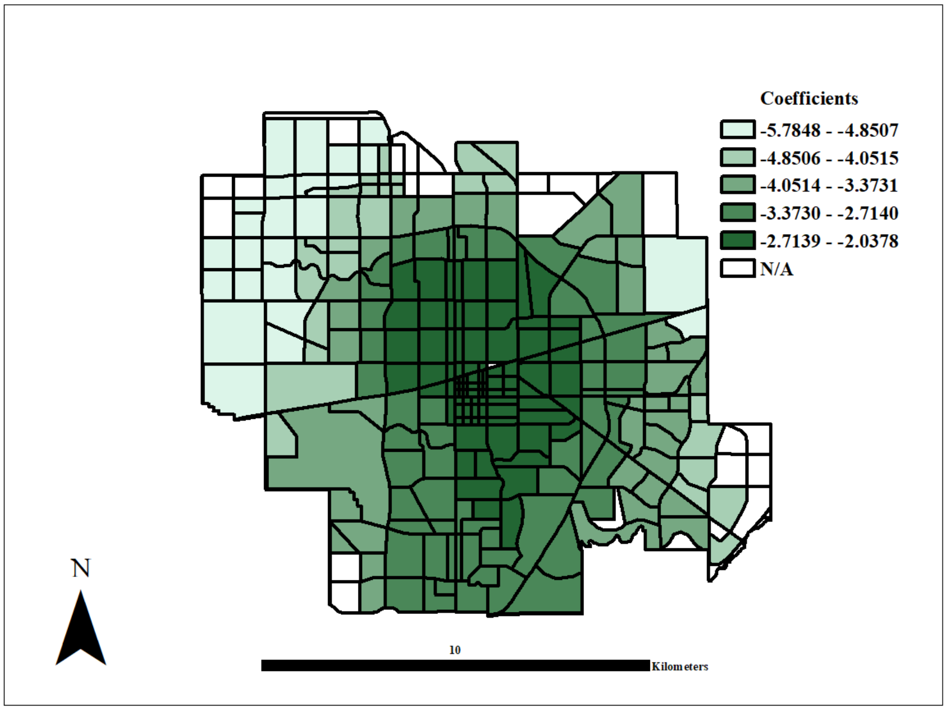
a) Fixed Gaussian Bandwidth GWPR



b) Fixed Gaussian Bandwidth GWNBR

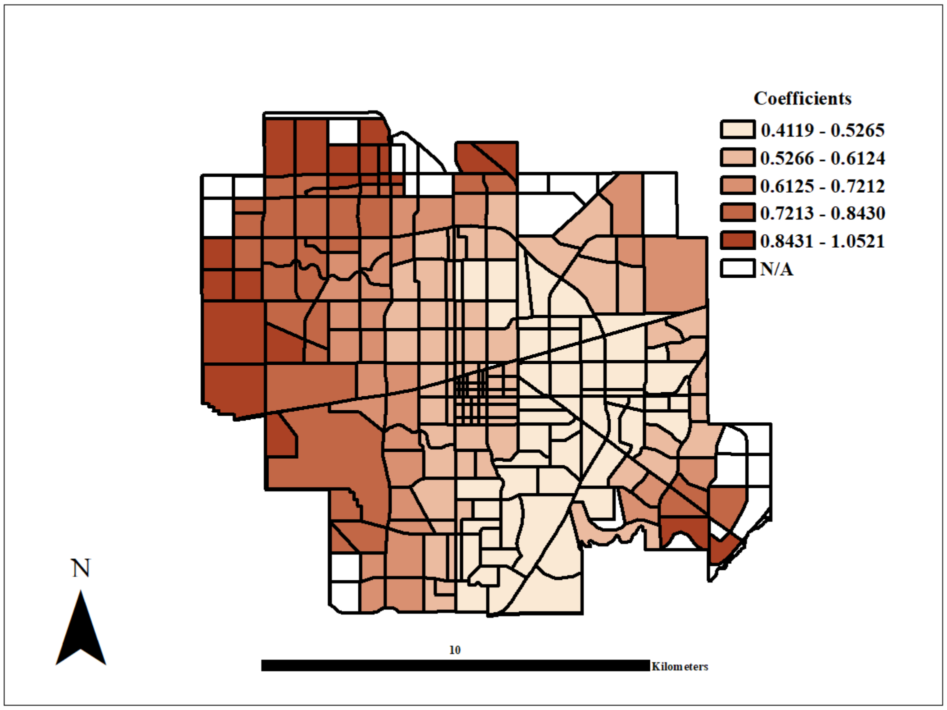


c) Adaptive Bi-square Bandwidth GWPR

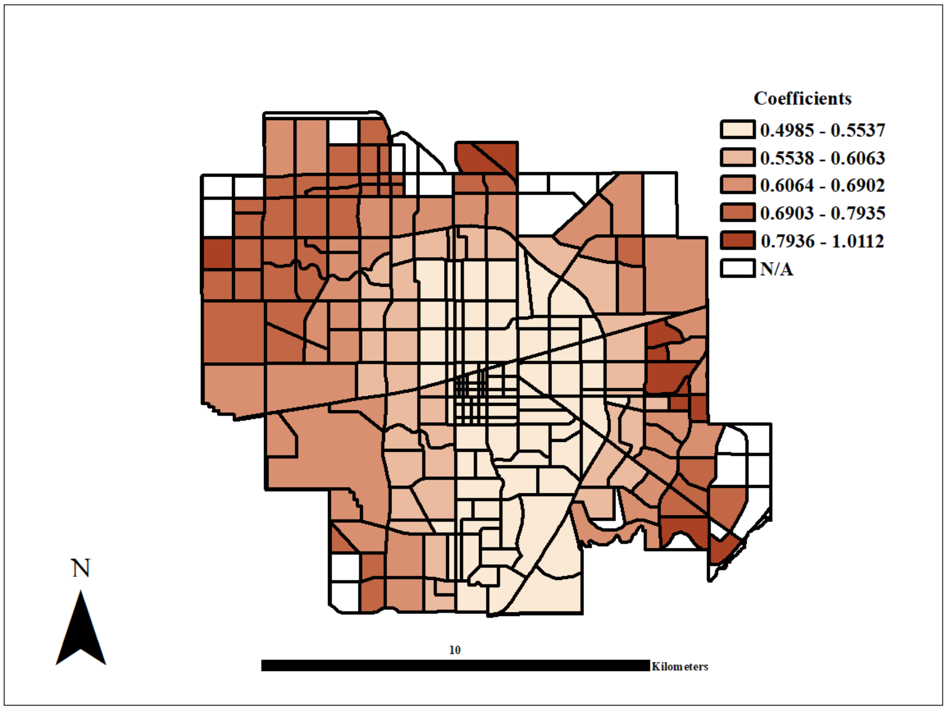


d) Adaptive Bi-square Bandwidth GWNBR

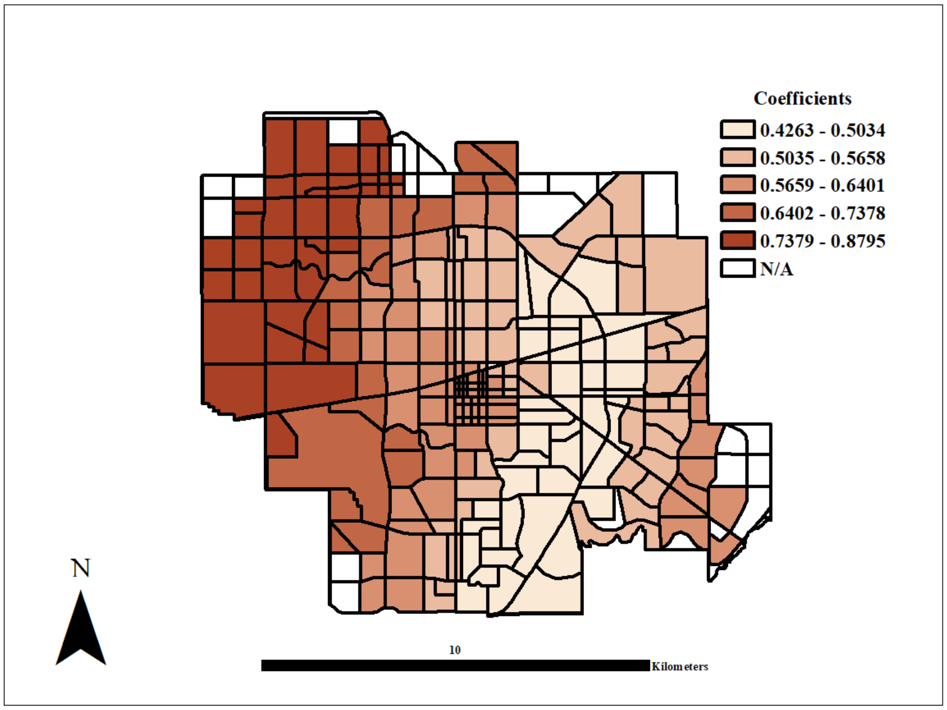
B. LOG TAZ VKMT



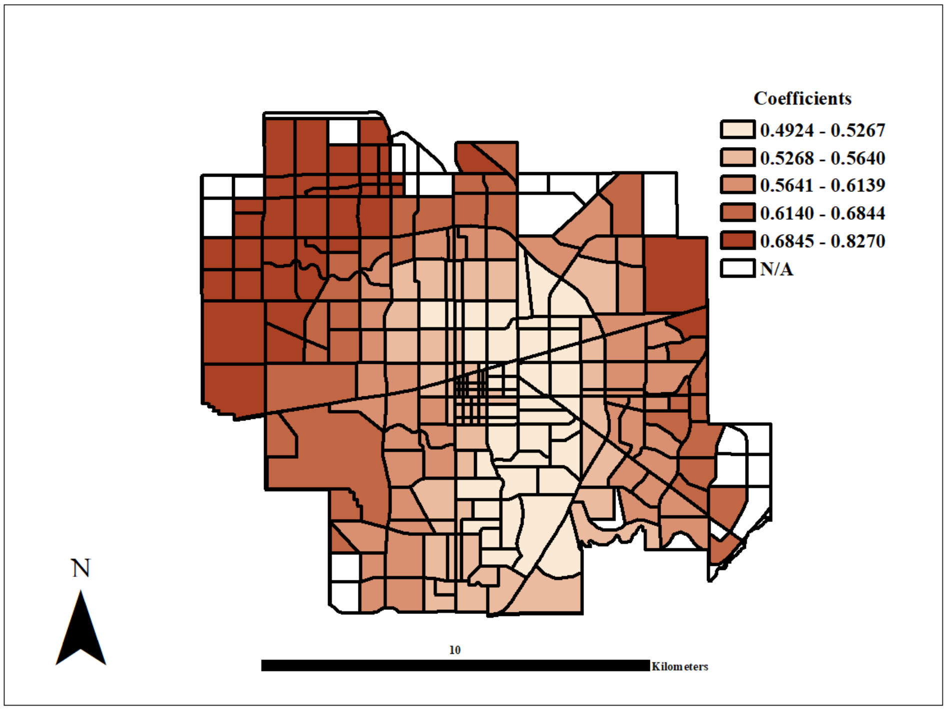
a) Fixed Gaussian Bandwidth GWPR



b) Fixed Gaussian Bandwidth GWNBR

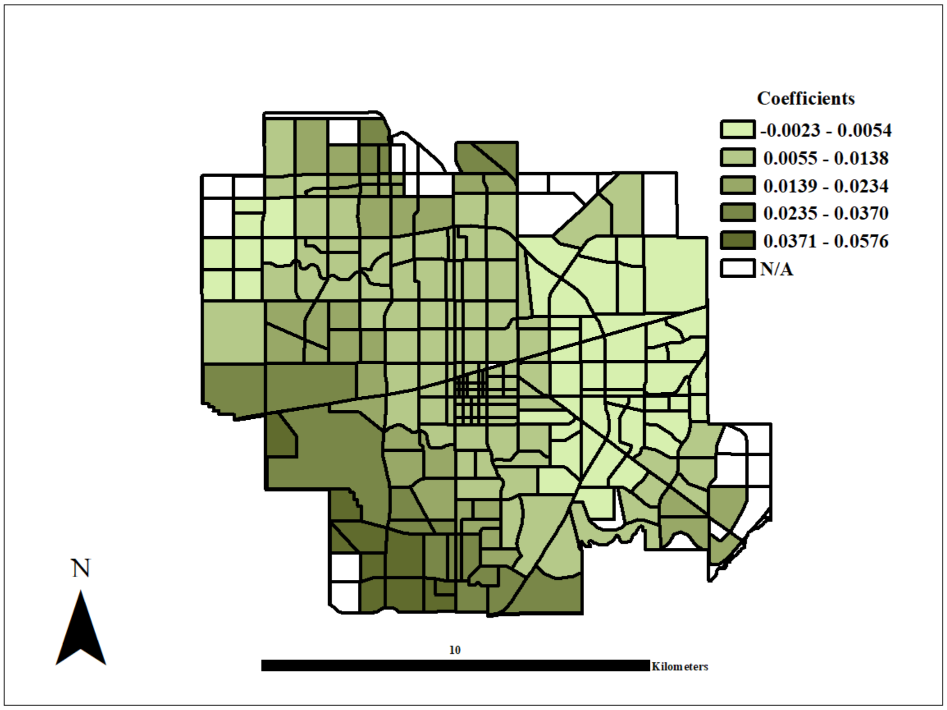


a) Adaptive Bi-square Bandwidth GWPR

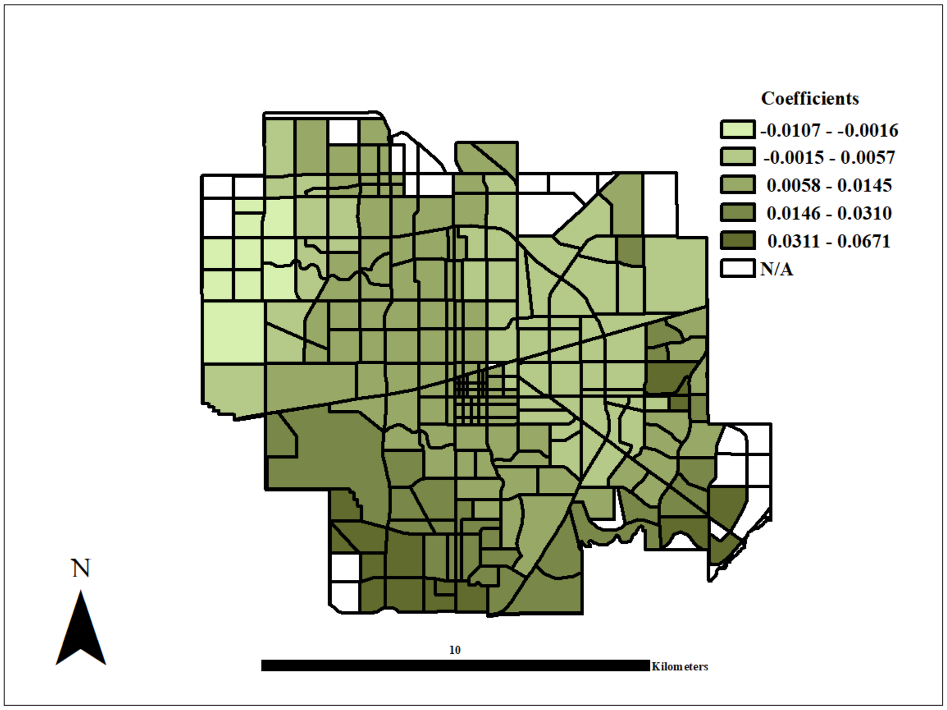


b) Adaptive Bi-square Bandwidth GWNBR

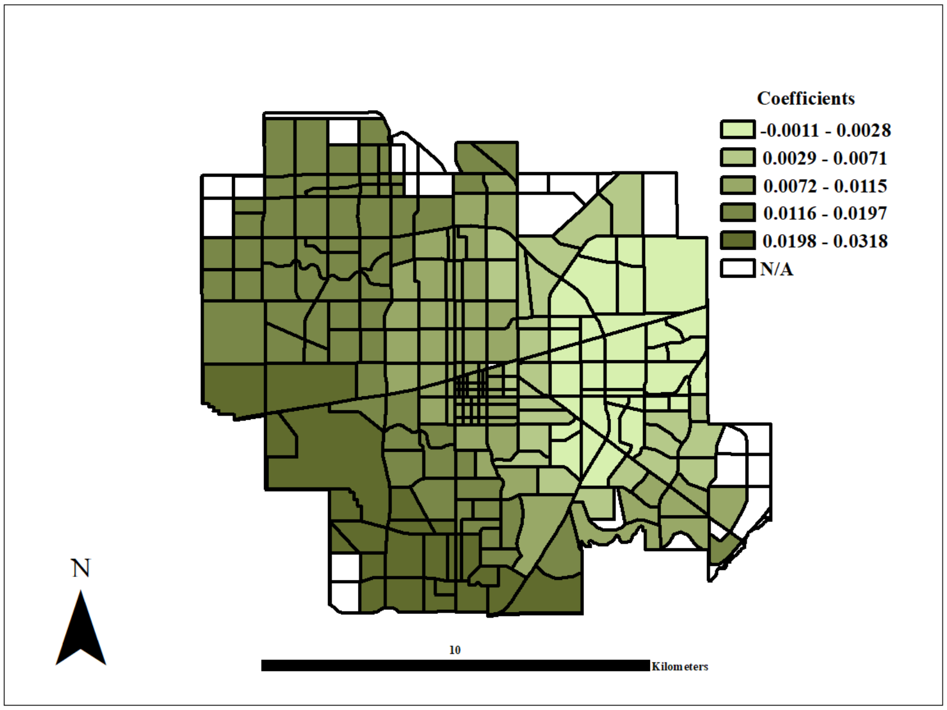
C. INTERSECTION DENSITY



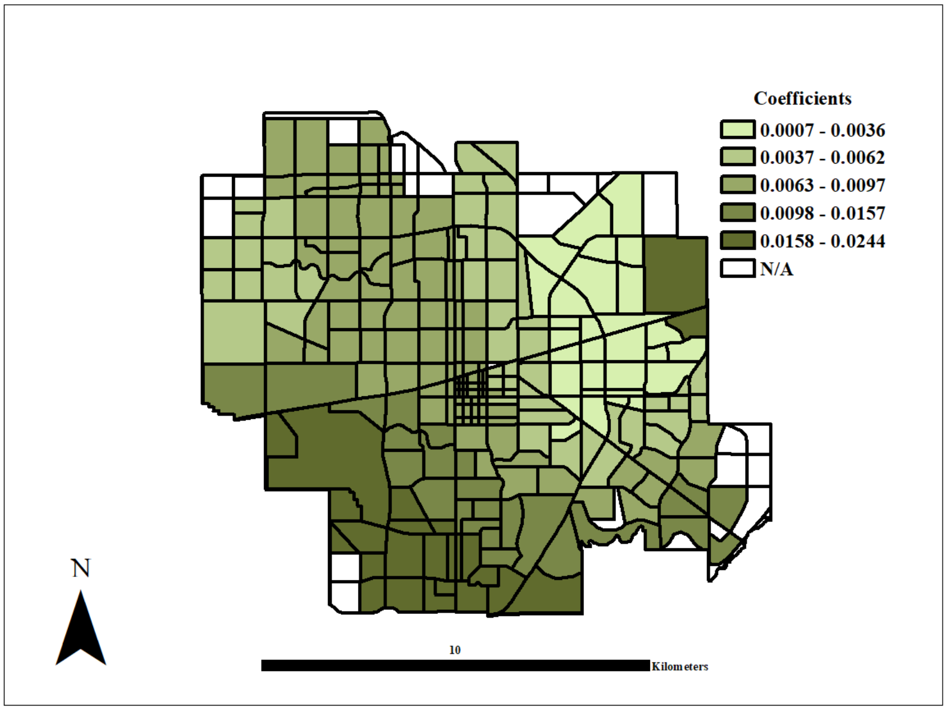
a) Fixed Gaussian Bandwidth GWPR



b) Fixed Gaussian Bandwidth GWNBR

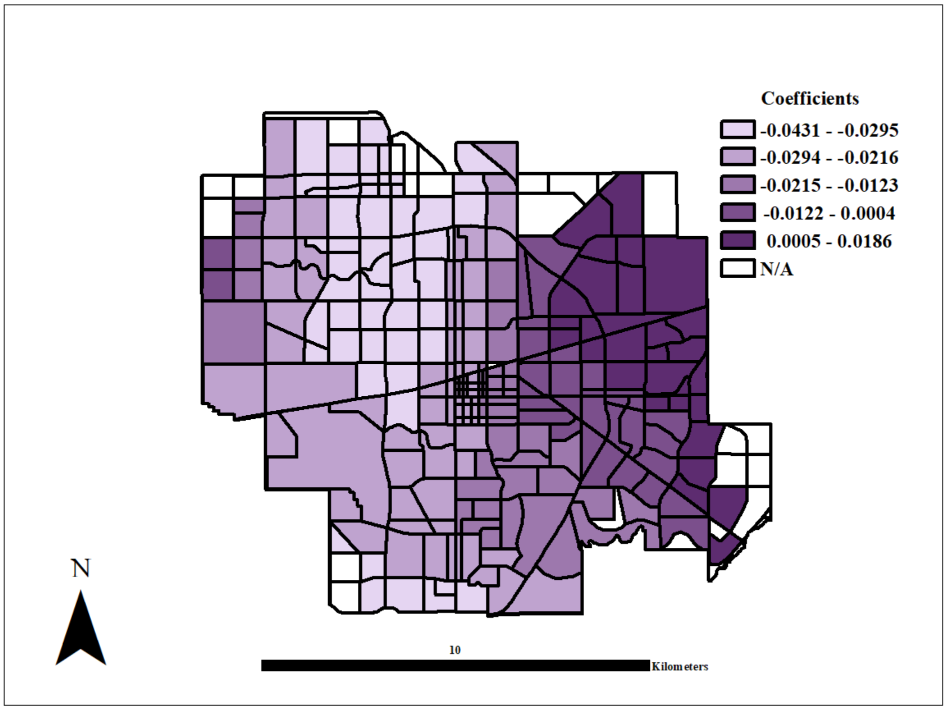


c) **Adaptive Bi-square Bandwidth GWPR**

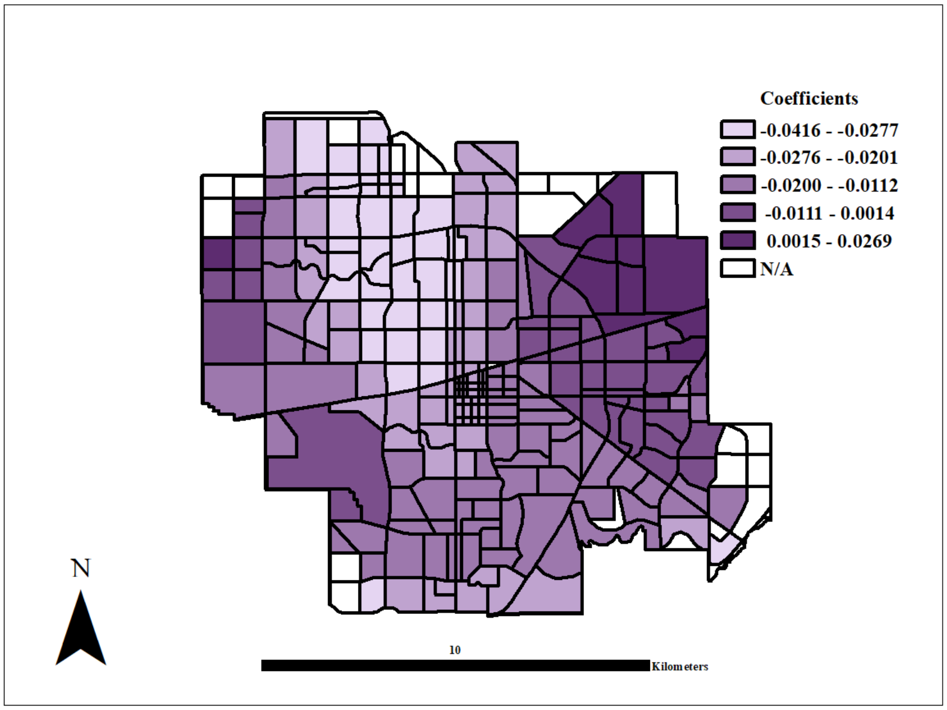


d) **Adaptive Bi-square Bandwidth GWNBR**

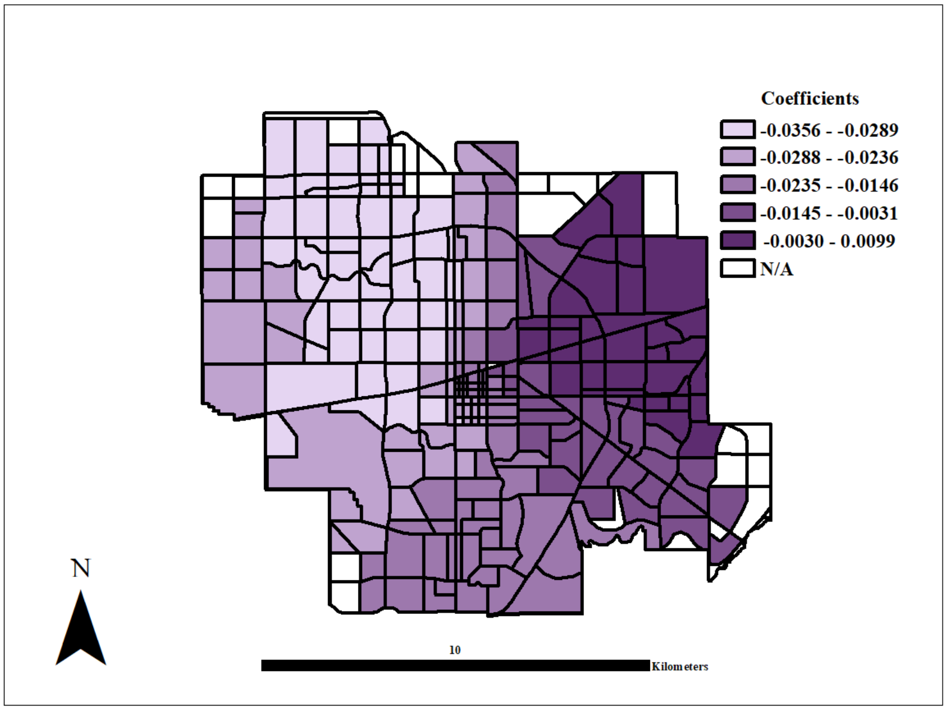
D. NUMBER OF 3 LEG INTERSECTIONS



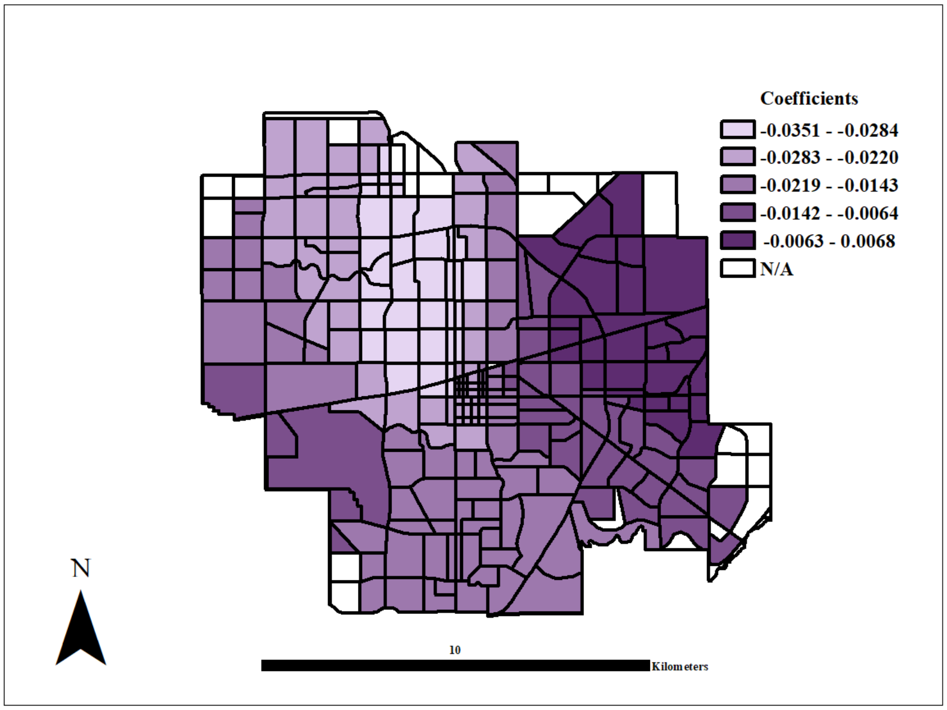
a) Fixed Gaussian Bandwidth GWPR



b) Fixed Gaussian Bandwidth GWNBR

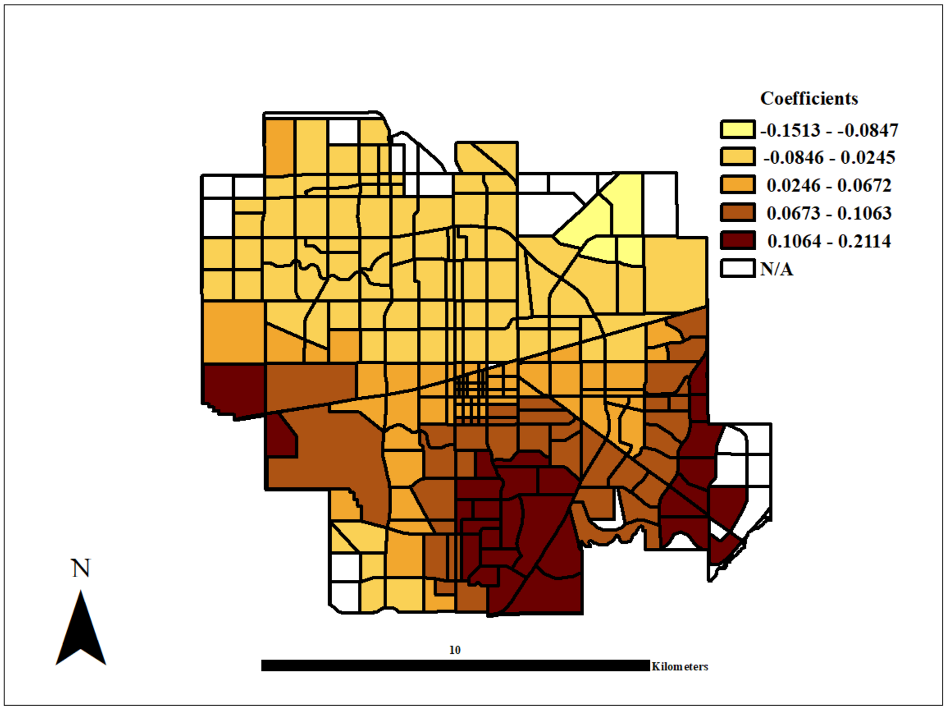


c) **Adaptive Bi-square Bandwidth GWPR**

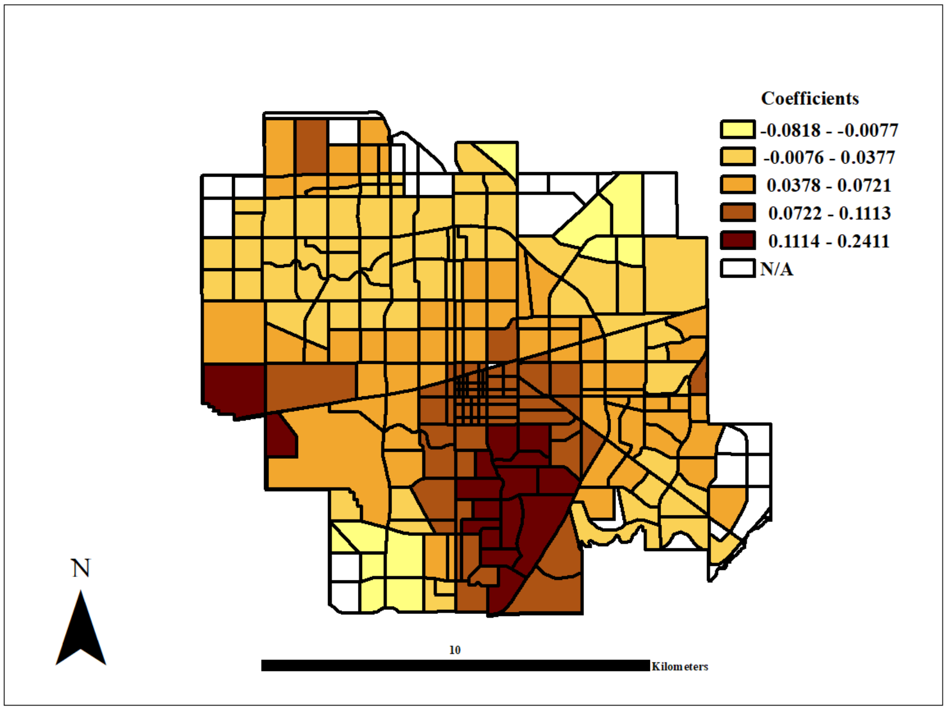


d) **Adaptive Bi-square Bandwidth GWNBR**

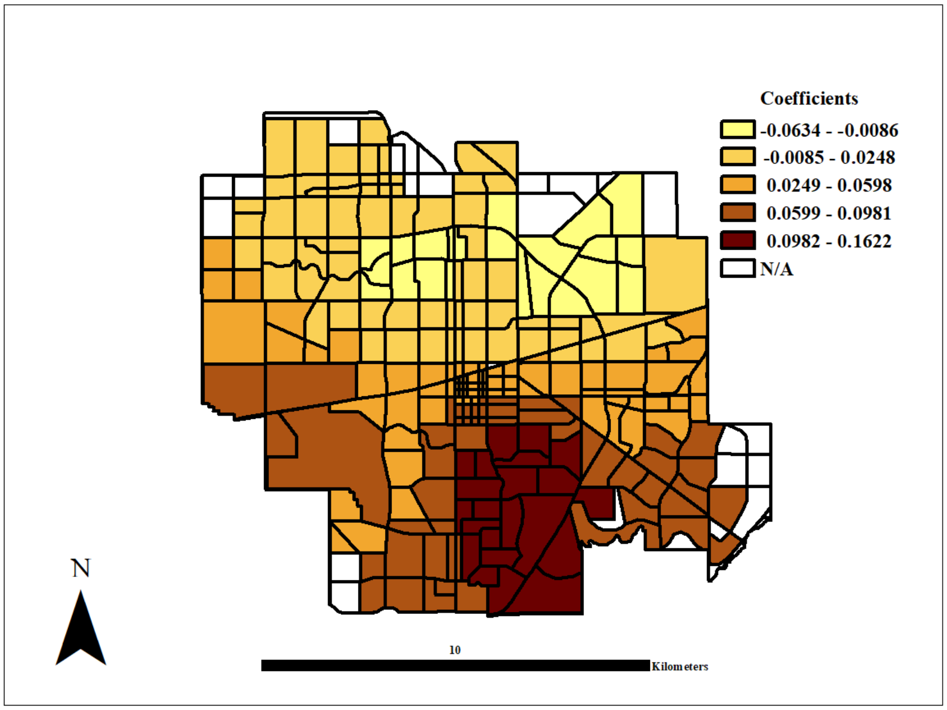
E. NUMBER OF LAND USES PER TAZ



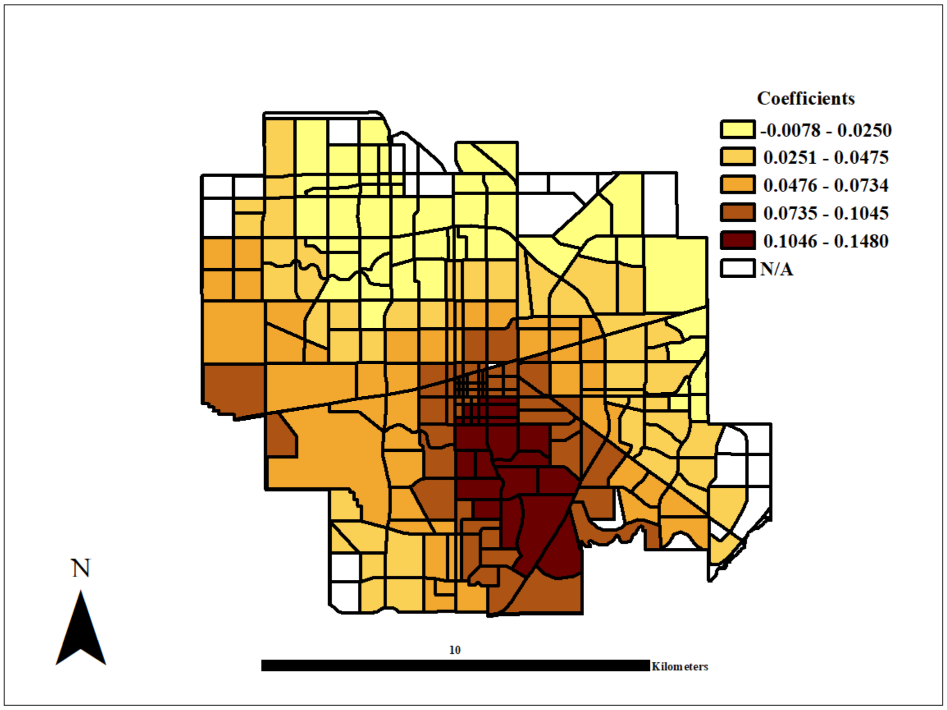
a) Fixed Gaussian Bandwidth GWPR



b) Fixed Gaussian Bandwidth GWNBR

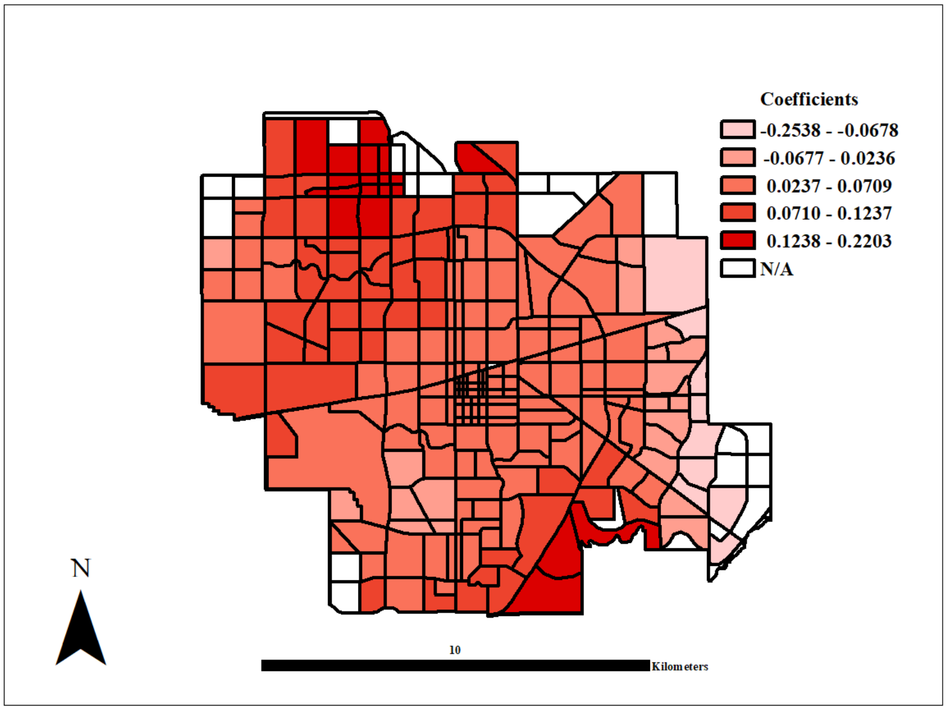


c) **Adaptive Bi-square Bandwidth GWPR**

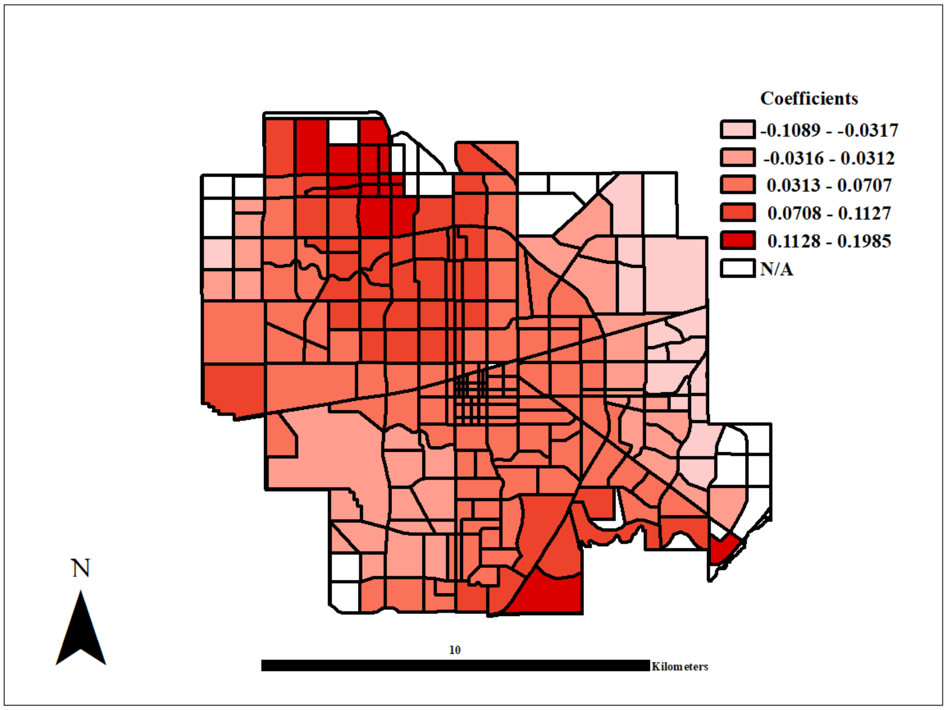


d) **Adaptive Bi-square Bandwidth GWNBR**

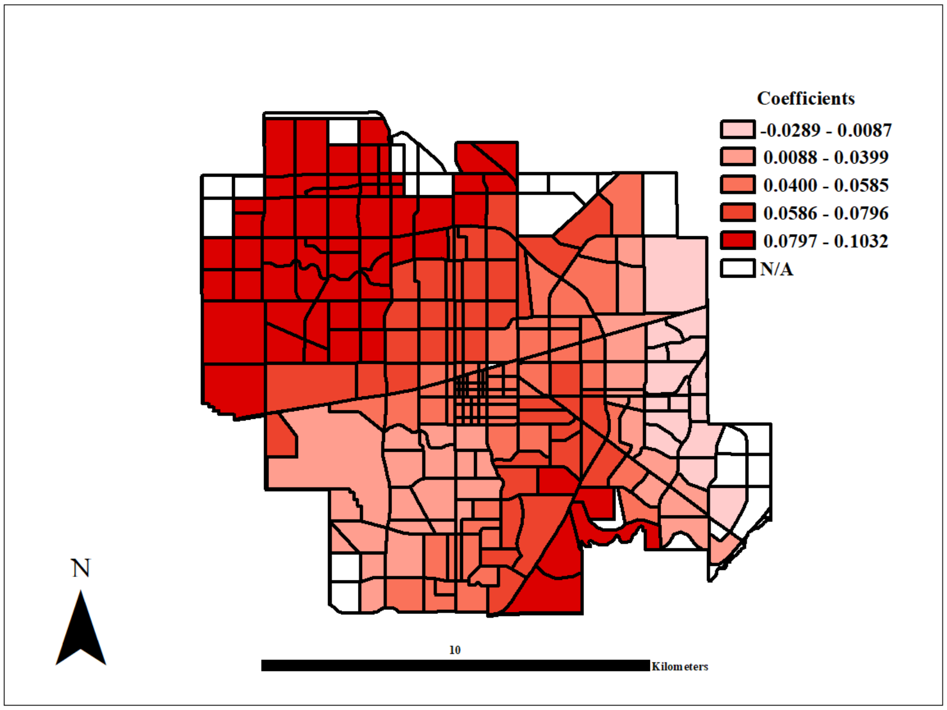
F. WEIGHTED AVERAGE SEGMENT LENGTH



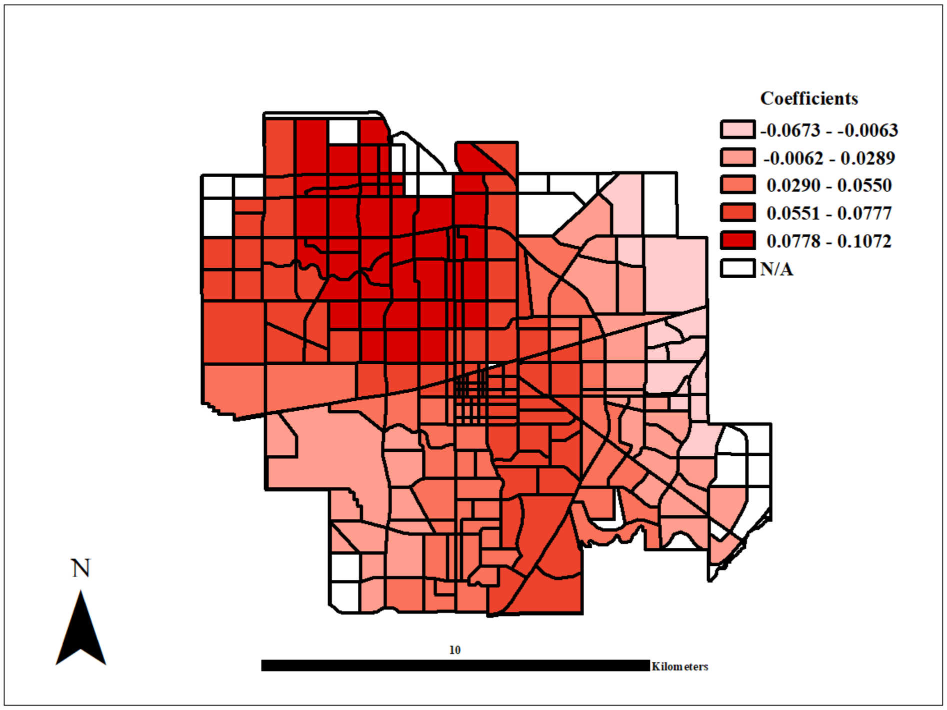
a) Fixed Gaussian Bandwidth GWPR



b) Fixed Gaussian Bandwidth GWNBR

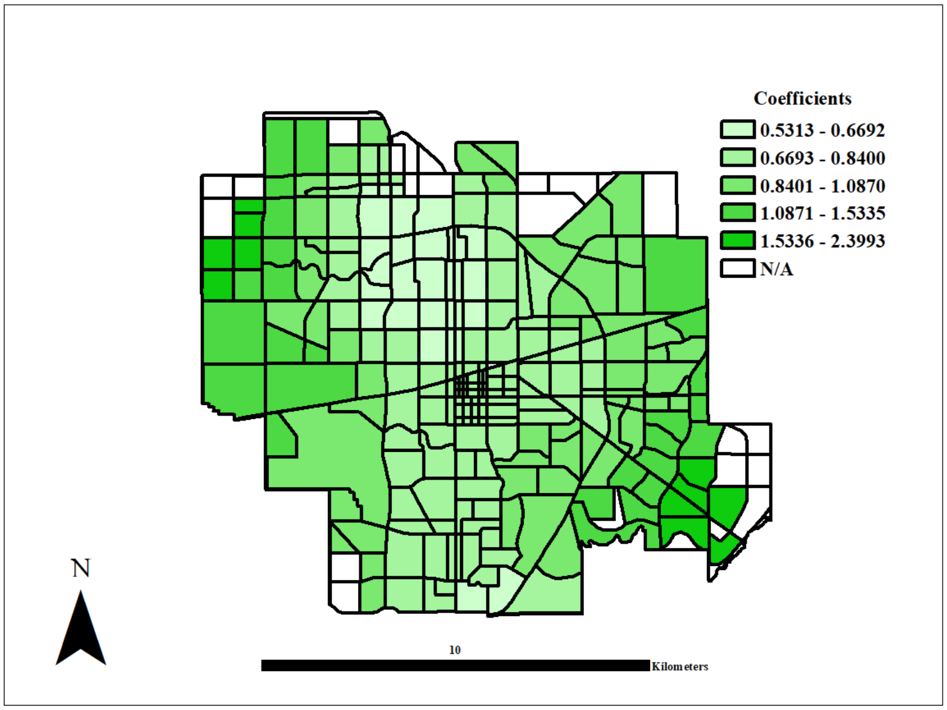


c) **Adaptive Bi-square Bandwidth GWPR**

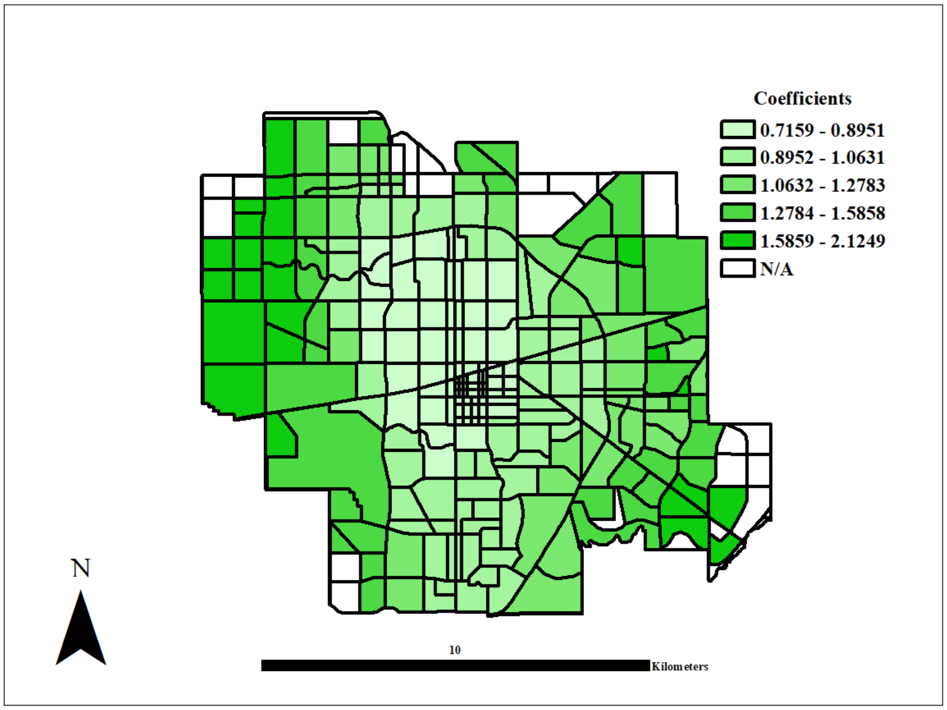


d) **Adaptive Bi-square Bandwidth GWPR**

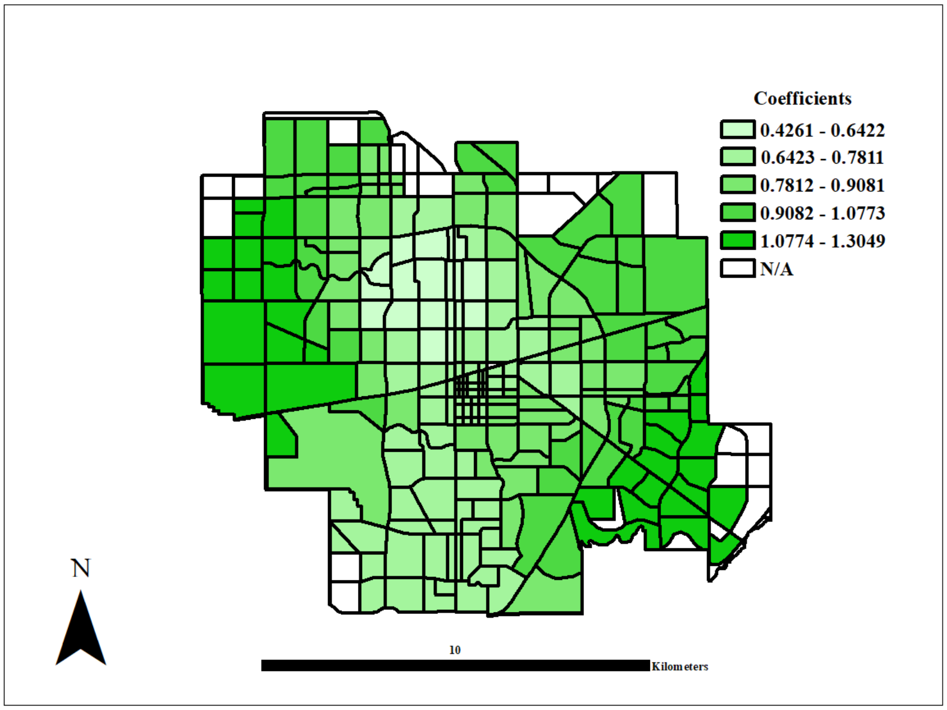
G. COMMERCIAL AREA



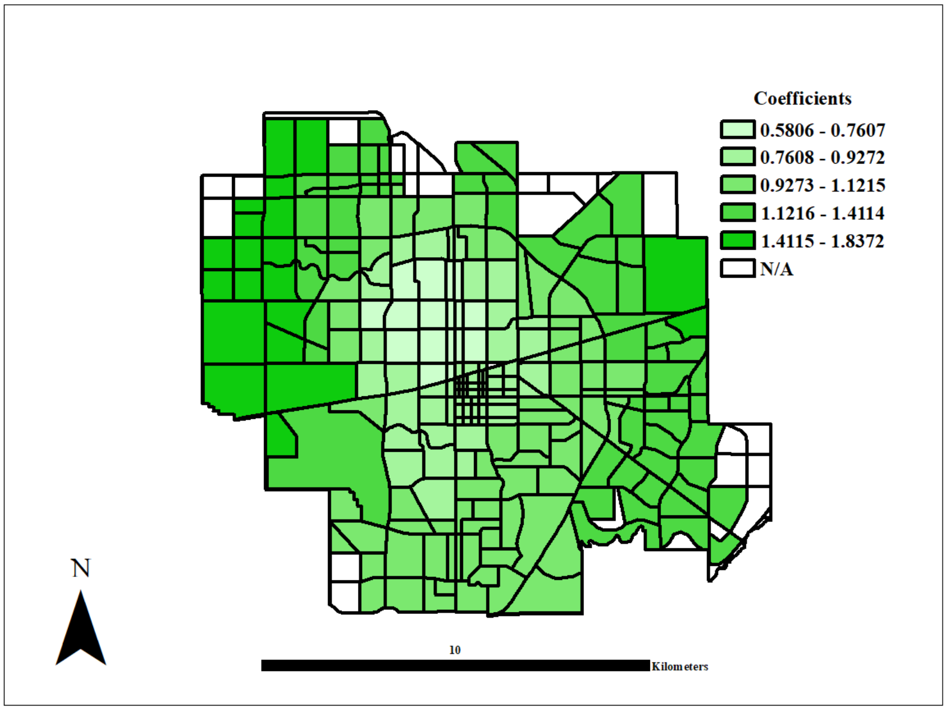
a) Fixed Gaussian Bandwidth GWPR



b) Fixed Gaussian Bandwidth GWNBR

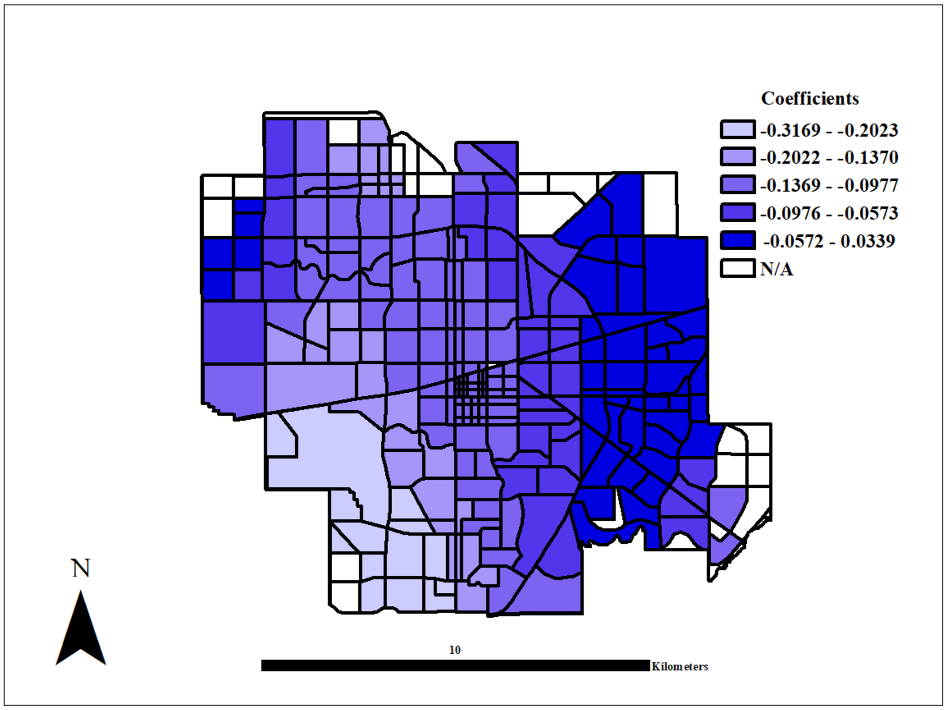


c) Adaptive Bi-square Bandwidth GWPR

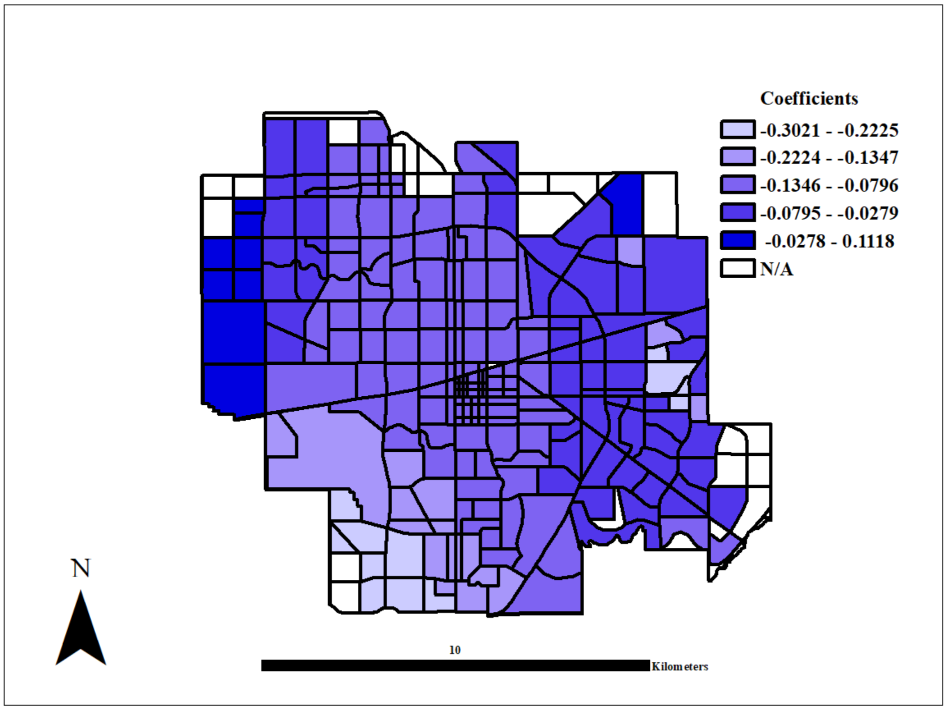


d) Adaptive Bi-square Bandwidth GWNBR

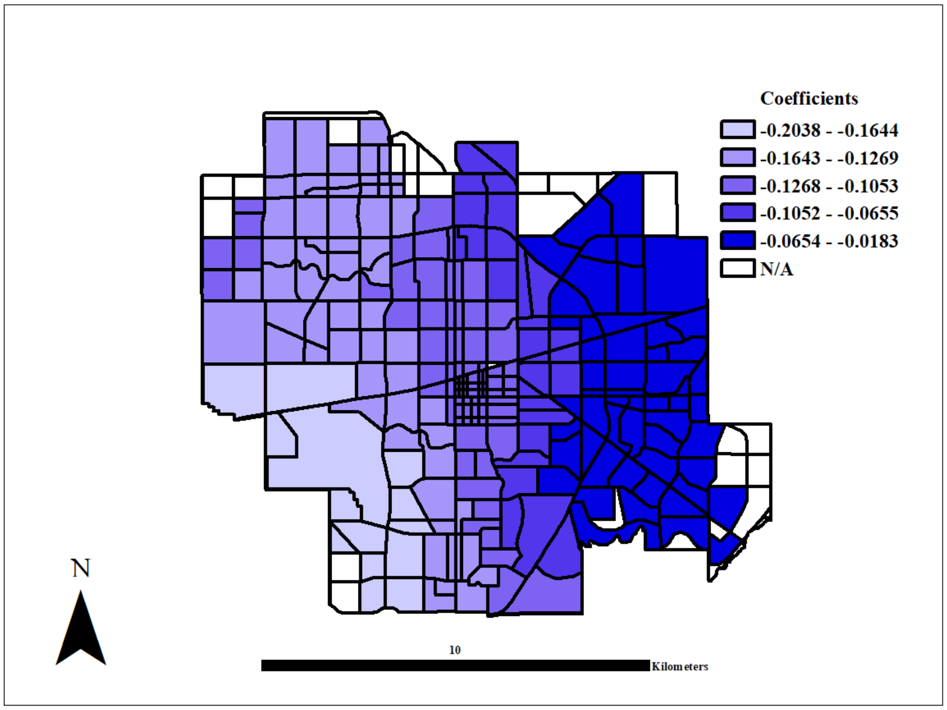
H. TOTAL SEGMENT LENGTH DENSITY



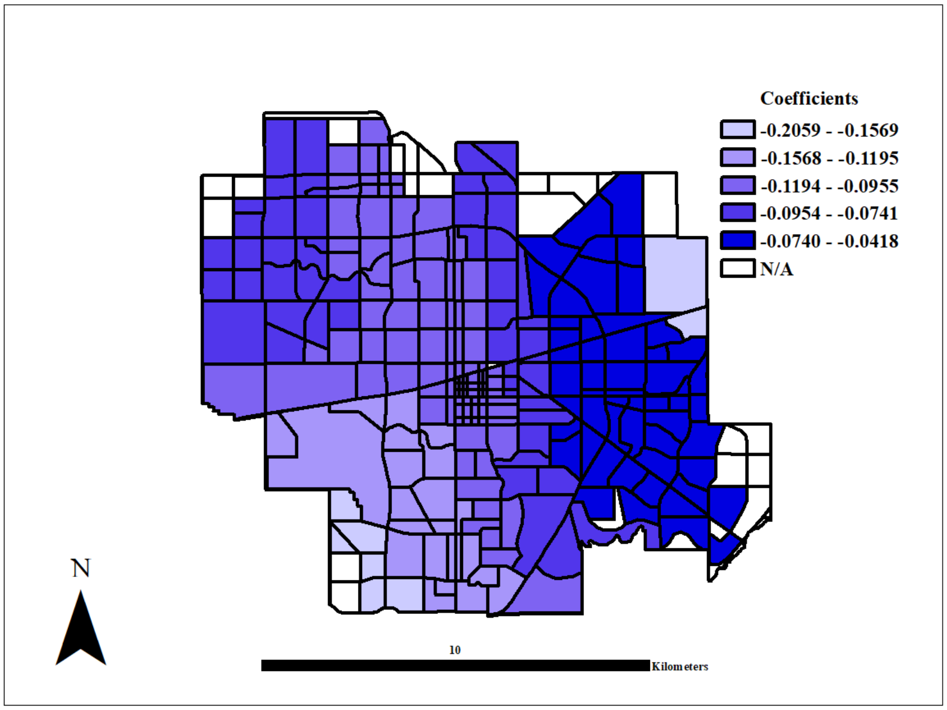
a) Fixed Gaussian Bandwidth GWPR



b) Fixed Gaussian Bandwidth GWNBR

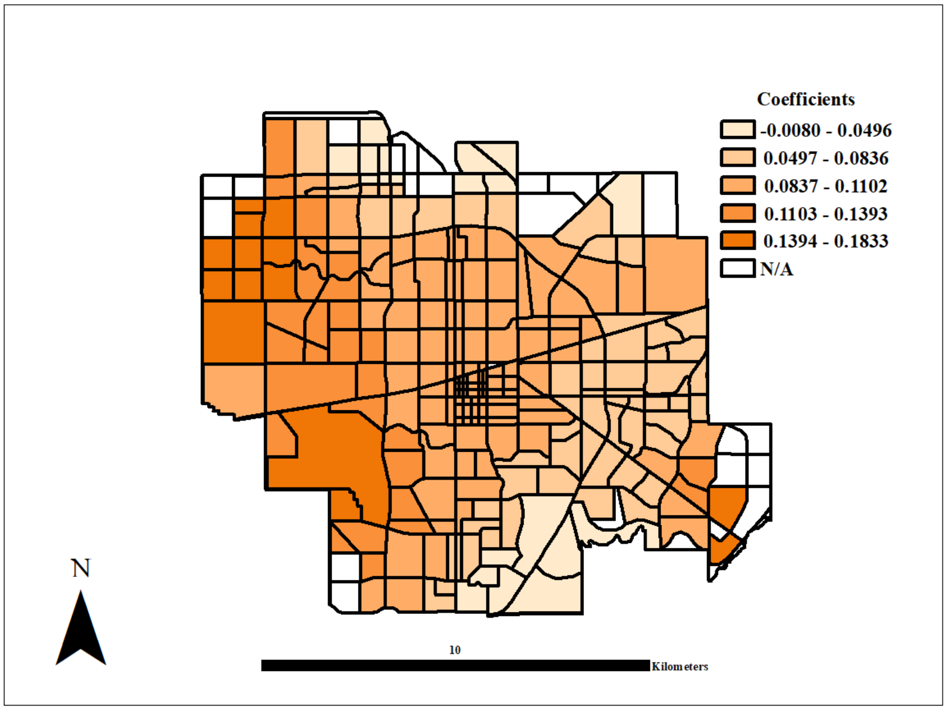


c) Adaptive Bi-square Bandwidth GWPR

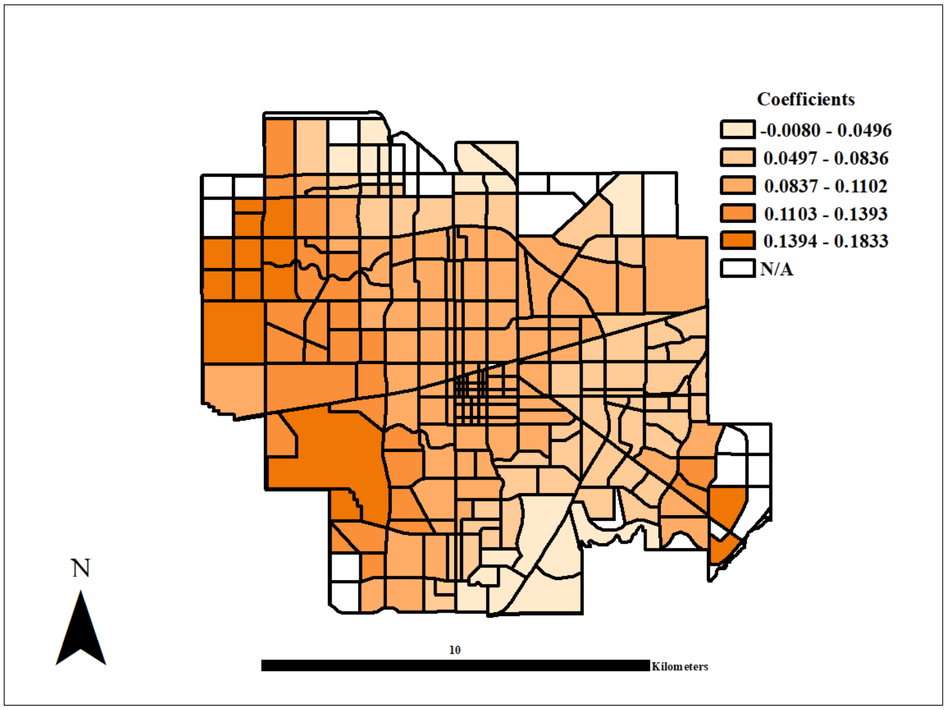


d) Adaptive Bi-square Bandwidth GWNBR

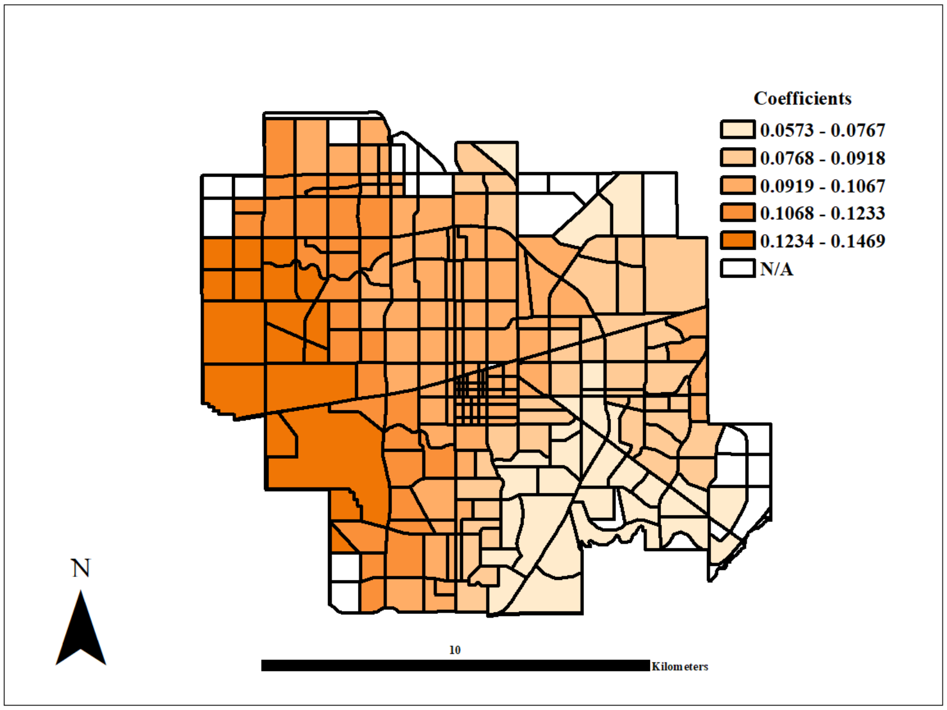
I. LOCAL ROAD DENSITY



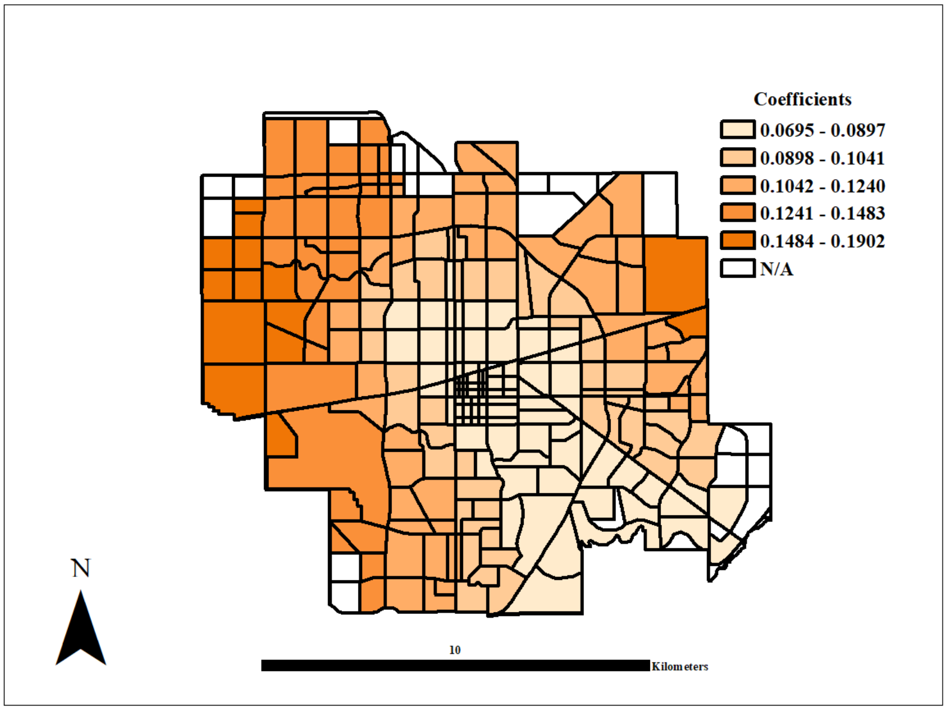
a) **Fixed Gaussian Bandwidth GWPR**



b) **Fixed Gaussian Bandwidth GWNBR**



c) **Adaptive Bi-square Bandwidth GWPR**



d) **Adaptive Bi-square Bandwidth GWNBR**

Appendix G: Negative Binomial (NB) Model for Total, Violent and Non-violent Crimes

Table G-1: Functional Forms for Best Predicting Models for Total, Violent and Non-Violent Crimes.

Severity	Functional Form
Total Crimes	$\bar{Y} = exp^{\beta_0+\beta_1 \ln(TOP\_POP)+\beta_2\times\%POP\_65plus+\beta_3\times\%RETAIL\_SPACE+\beta_4\times\%OFFICE\_SPACE+\beta_5\times\%URBAN\_HOLDING\_AREA+\beta_6\times TAZ\_AREA}$
Violent Crimes	$\bar{Y} = exp^{\beta_0+\beta_1 \ln(TOP\_POP)+\beta_2\times\%INDUSTRY\_SPACE+\beta_3\times NO\_LU\_PER\_TAZ+\beta_4\times\%COMMERCIAL\_AREA+\beta_5\times\%RESIDENTIAL\_MD\_+\beta_6\times\%URBAN\_HOLDING\_AREA}$
Non-Violent Crimes	$\bar{Y} = exp^{\beta_0+\beta_1 \ln(TOP\_POP)+\beta_2\times\%POP_{65plus}+\beta_3\times\%RETAIL_{SPACE}+\beta_4\times\%INDUSTRY\_SPACE+\beta_5\times NO\_LU\_PER\_TAZ+\beta_6\times\%COMMERCIAL\_AREA+\beta_7\times\%URBAN\_HOLDING\_AREA}$

Table G-2: Model Parameters for Total, Violent and Non-Violent Crimes

Parameter	Total Crimes			Violent Crimes			Non-Violent Crimes		
	Estimate	Standard Error	Pr > ChiSq	Estimate	Standard Error	Pr > ChiSq	Estimate	Standard Error	Pr > ChiSq
Intercept	2.3407	0.2056	<.0001	-1.9046	0.4022	<.0001	0.7811	0.2637	0.0031
LOG_TOT_POP	0.2630	0.0272	<.0001	0.2816	0.0529	<.0001	0.2421	0.0322	<.0001
POP_65plus	-2.8738	0.5359	<.0001				-1.2738	0.5166	0.0137
INDUSTRY_SPACE	-	-	-	11.5335	3.6572	0.0016	8.6977	2.4323	0.0003
RETAIL_SPACE	3.7418	1.0265	0.0003	-	-	-	3.0573	1.0022	0.0023
OFFICE_SPACE	0.6221	0.3359	0.064	-	-	-	-	-	-
NO_LU_PER_TAZ	-	-	-	0.3838	0.0752	<.0001	0.2574	0.0477	<.0001
COMMERCIAL_AREA	-	-	-	1.2134	0.3247	0.0002	0.7030	0.2769	0.0111
RESIDENTIAL_MD_AREA	-	-	-	2.1815	0.6812	0.0014	-	-	-
URBAN_HOLDING_AREA	-3.999	0.3741	<.0001	-3.7623	0.569	<.0001	-3.0625	0.2925	<.0001
TAZ_AREA	0.3125	0.1535	0.0418	-	-	-	-	-	-
Dispersion		0.6980			1.0970			0.5469	

Table G-3: Estimated Parameters for Geographically Weighted Poisson (GWPR) Model

Crimes Category	Variables	Adaptive Bandwidth				
		Min	1st Quart	Median	3rd Quart	Max
Total Crimes	Intercept	-1.218	1.206	2.228	2.601	3.178
	LOG_TOT_POP	0.069	0.239	0.266	0.425	0.664
	POP_65plus	-7.329	-4.326	-2.421	-0.721	2.456
	RETAIL_SPACE	-0.453	2.496	2.916	3.313	6.357
	OFFICE_SPACE	-1.411	0.924	1.109	1.279	10.435
	URBAN_HOLDING_AREA	-98.775	-8.289	-4.079	-2.946	10.669
	TAZ_AREA	-0.324	0.169	0.534	1.099	2.330
Violent Crimes	Intercept	-2.057	-1.754	-1.608	-1.474	-1.124
	LOG_TOT_POP	0.360	0.394	0.416	0.448	0.475
	INDUSTRY_SPACE	5.962	7.440	9.601	11.224	15.309
	NO_LU_PER_TAZ	0.173	0.195	0.208	0.220	0.263
	COMMERCIAL_AREA	0.424	0.570	0.639	0.718	0.824
	RESIDENTIAL_MD_AREA	1.069	1.304	1.445	1.628	1.914
	URBAN_HOLDING_AREA	-6.536	-6.220	-5.988	-5.284	-4.257
Non-Violent Crimes	Intercept	0.690	0.953	1.122	1.259	1.445
	LOG_TOT_POP	0.223	0.251	0.278	0.302	0.339
	POP_65plus	-2.449	-2.247	-1.940	-1.570	-0.789
	RETAIL_SPACE	2.459	3.018	3.274	3.418	3.643
	INDUSTRY_SPACE	5.823	7.219	8.112	9.155	11.552
	NO_LU_PER_TAZ	0.158	0.191	0.201	0.228	0.243
	COMMERCIAL_AREA	0.062	0.180	0.206	0.271	0.533
	URBAN_HOLDING_AREA	-7.372	-5.142	-4.707	-4.231	-3.368

Table G-4: Estimated Parameters for Geographically Weighted Negative Binomial (GWNBR) Model

Crimes Category	Parameter	Adaptive Bandwidth				
		Min	1st Quant	Median	3rd Quant	Max
Total Crimes	**Intercept**	-62392901.8200	2.4737	2.5362	2.6261	90099797.7040
	LOG_TOT_POP	-7851536.1500	0.2432	0.2514	0.2529	26793784.5120
	POP_65plus	-298163996.9000	-3.4281	-3.1318	-2.8188	-2.3730
	RETAIL_SPACE	-186616550.8000	3.1493	3.2264	3.3414	238446441.5600
	OFFICE_SPACE	-679355.8551	0.4918	0.5682	0.6190	84894147.2760
	URBAN_HOLDING_AREA	-92220346.7000	-4.6363	-4.5394	-4.3746	-4.1690
	TAZ_AREA	-4703472.2390	0.2847	0.3847	0.5369	29647866.7450
Violent Crimes	**Intercept**	-94328005.4800	-1.7140	-1.1026	-0.6014	112891935.3500
	LOG_TOT_POP	-1951723.6380	0.2001	0.3087	0.3669	8223417.0085
	INDUSTRY_SPACE	-908444169.6000	4.8219	9.3413	15.1227	5883035614.9000
	NO_LU_PER_TAZ	-15278109.9300	0.1972	0.2718	0.4456	23525793.6150
	COMMERCIAL_AREA	-97559746.3000	0.0829	0.3748	0.8143	46601524.6990
	RESIDENTIAL_MD_AREA	-276492811.7000	0.7392	1.7318	2.6945	73041630.7750
	URBAN_HOLDING_AREA	-128978043.4000	-8.9133	-4.0457	-2.3187	27178667.9620
Non-Violent Crimes	**Intercept**	-100855625.9000	0.9084	1.0210	1.1496	60558203.0080
	LOG_TOT_POP	-5456953.5850	0.2320	0.2499	0.2595	17831606.4270
	POP_65plus	-252691816.6000	-1.6885	-1.4815	-1.1609	121577855.4200
	RETAIL_SPACE	-89815764.9700	2.9032	3.1038	3.1777	402820506.8200
	INDUSTRY_SPACE	-344566897.7000	7.7809	8.4531	9.3529	556706394.4100
	NO_LU_PER_TAZ	-7793969.7700	0.2176	0.2213	0.2574	29368313.0370
	COMMERCIAL_AREA	-45023473.3700	0.4419	0.4809	0.5533	56611803.8930
	URBAN_HOLDING_AREA	-66848173.8600	-4.2961	-3.5304	-3.2916	37285984.8720

Table G-5: Crime Types and Sub-types

General Types	Crime Sub-types
Assault	Aggravated Assault-Level 3
	Assault-Common Level 1
	Assault-Other
	CC-Criminal Negligence causing bodily harm
	Assault with weapon or cause bodily Harm Level 2
	Discharge Firearm with intent
	DVC Aggravated Assault-Level 3
	DVC Assault-Common Level 1
	DVC Assault with weapon or cause bodily Harm-Level 2
	Pointing a firearm
	Using firearms (or imitation) in commission of offence
Arson	Arson
	Arson-Disregard for Human Life
Break and Enter	Break and Enter
	Break and Enter-Home Invasion
	Break and Enter-Firearms
	Break and Enter-Compound
	Break and Enter to Motor Vehicle-Firearms
	DVC Break and Enter
Robbery	Robbery
	Robbery-Commercial
	Robbery-Delivery Person
	Robbery-Financial Institution
	Robbery-Purse-snatching
	Robbery-Street
	Robbery-Taxi
Sexual Assault	Aggravated Sexual Assault
	DVC Aggravated Sexual Assault
	Sexual Assault
	Sexual Assault with a weapon
	DVC Sexual Assault
	DVC Sexual Assault with a weapon
	Incest
	Invitation to sexual touching
	Luring a child via a computer
	Sexual exploitation
	Sexual exploitation of a person with a disability
	Sexual interference
	Sexually explicit material available to a child
	Voyeurism

Table G-5: Crime Types and Sub-types (Cont'd)

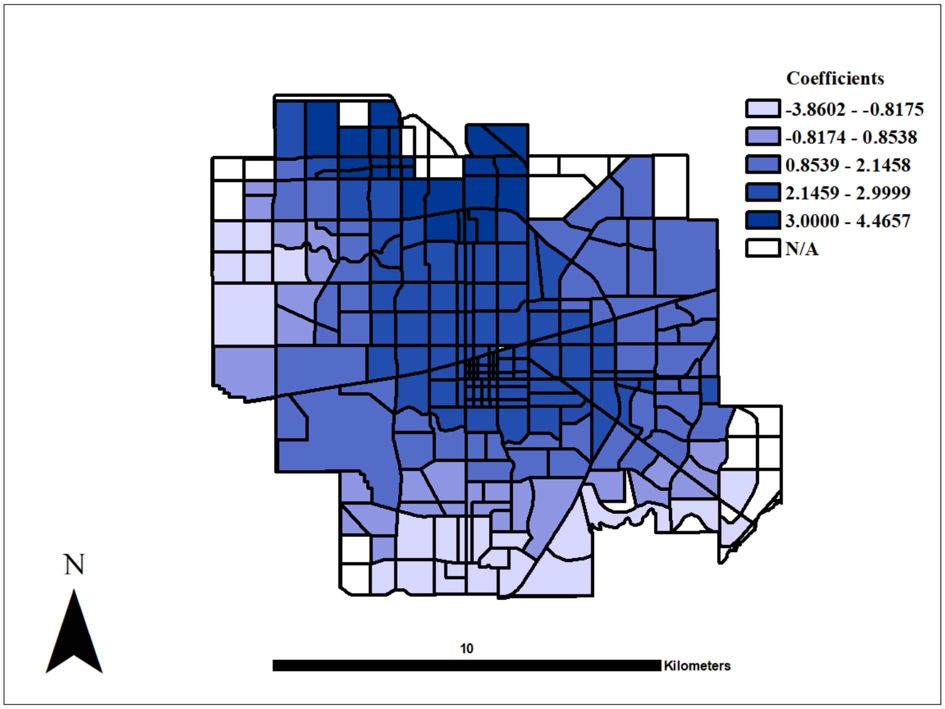
General Types	Crime Sub-types
Theft	DVC (Domestic Violence Court) Theft Over $5,000 DVC (Domestic Violence Court) Theft Under $5,000 Mail theft before delivery over $5,000 Mail theft before delivery under $5,000 Pick-pocketing over $5,000 Pick-pocketing under $5,000 Purse-snatching under $5,000 Shoplifting Over $5,000 Shoplifting $5,000 or under Theft Over $5,000 Theft Under $5,000 Theft of telecommunications over $5,000 Theft of telecommunications under $5,000
Mischief	Mischief-No damage Mischief Over $5,001 Mischief Under $5,000- Graffiti Mischief Under $5,001 Mischief Willful Damage Public Mischief DVC Mischief Over $5,001 DVC Mischief Under $5,001
Theft from Auto	Theft from Auto Over $5,000 Theft from Auto Under $5,000
Theft of Auto	Theft of Auto Theft of Auto Over $5,000 Theft of Auto Under $5,000

Table G-6: Moran I Test for Spatial Dependency in Crime Variable

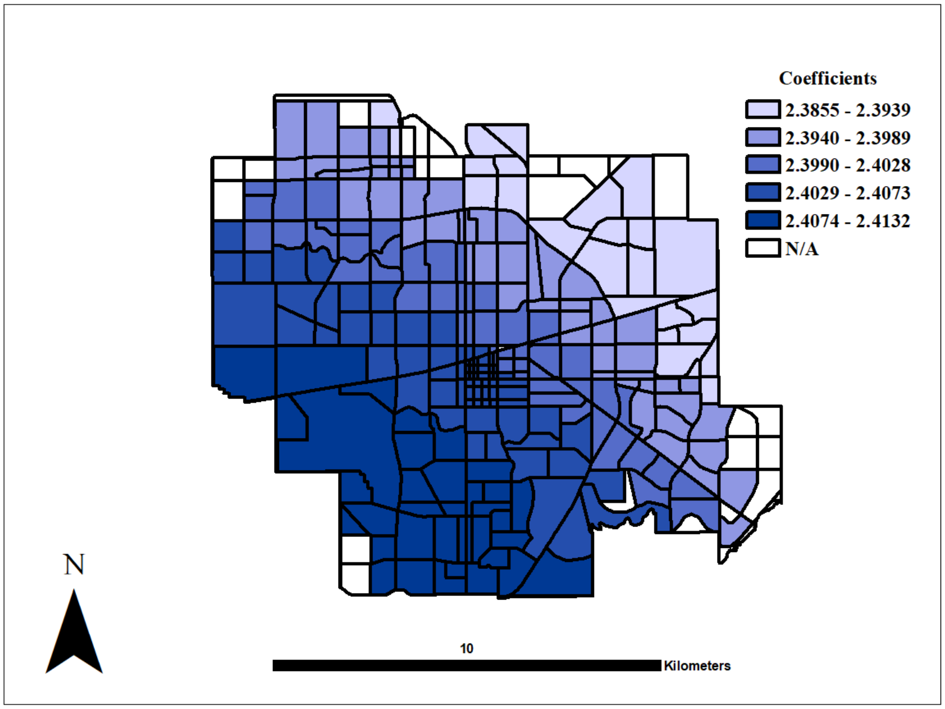
Variables	Moran I	Z	Pr > |Z|
Dependent			
TOTAL_CRIMES	0.0239	9.4300	<.0001
VIOLENT_CRIMES	0.0232	9.2100	<.0001
NON_VIOLENT_CRIME	0.0216	8.6600	<.0001
Independent			
TOT_POP	0.0421	15.5800	<.0001
POP_65plus	0.133	46.0900	<.0001
OFFICE_SPACE	0.111	38.8000	<.0001
URBAN_HOLDING_AREA	0.0311	11.9000	<.0001
TAZ_AREA	0.0476	17.4300	<.0001
RETAIL_SPACE	0.0345	13.0100	<.0001
INDUSTRY_SPACE	0.0368	13.7740	<.0001
NO_LU_PER_TAZ	0.0544	19.7034	<.0001
COMMERCIAL_AREA	0.0987	34.6400	<.0001
RESIDENTIAL_MD_AREA	0.0253	9.9100	<.0001

Appendix H: Thematic Representation of the Coefficients of GWPR and GWNBR Models for Total Crimes

A. INTERCEPT

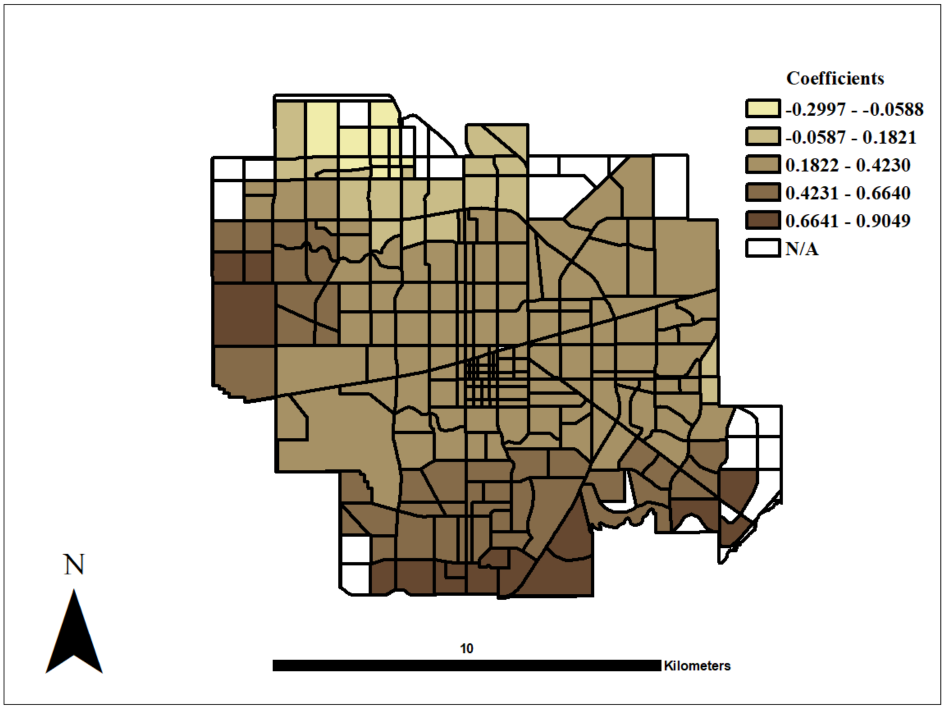


a) Fixed Gaussian Bandwidth GWPR

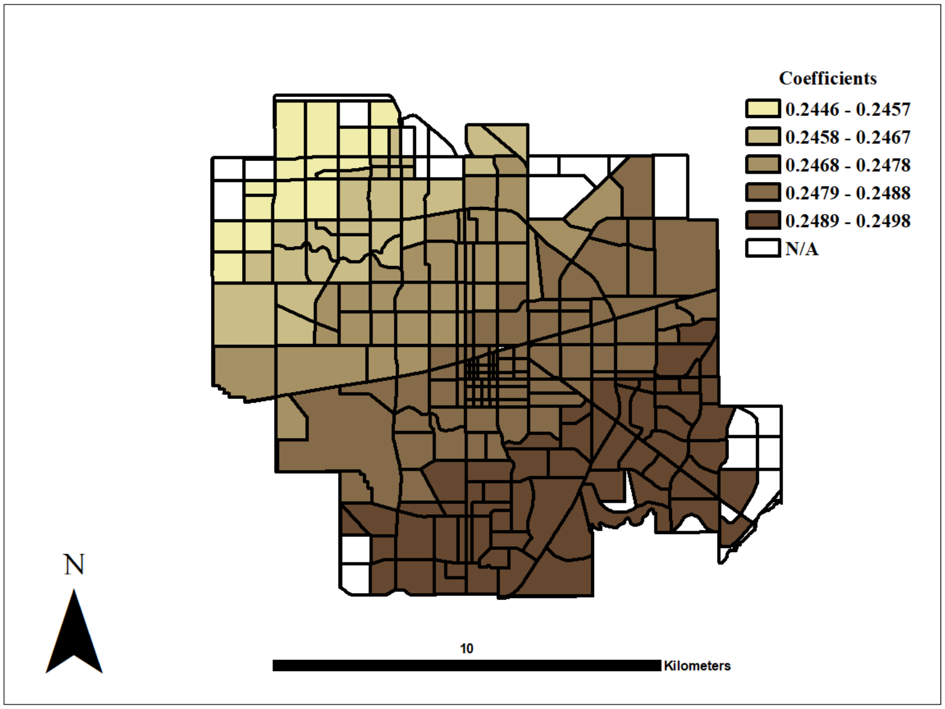


b) Fixed Gaussian Bandwidth GWNBR

B. LOG TOTAL POPULATION

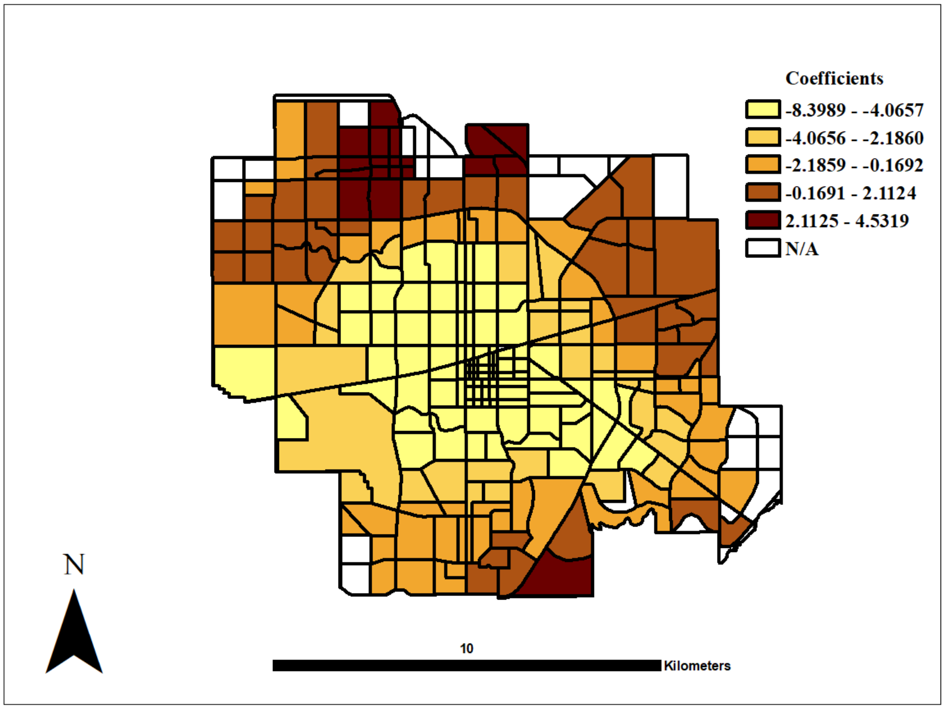


a) **Fixed Bandwidth GWPR**

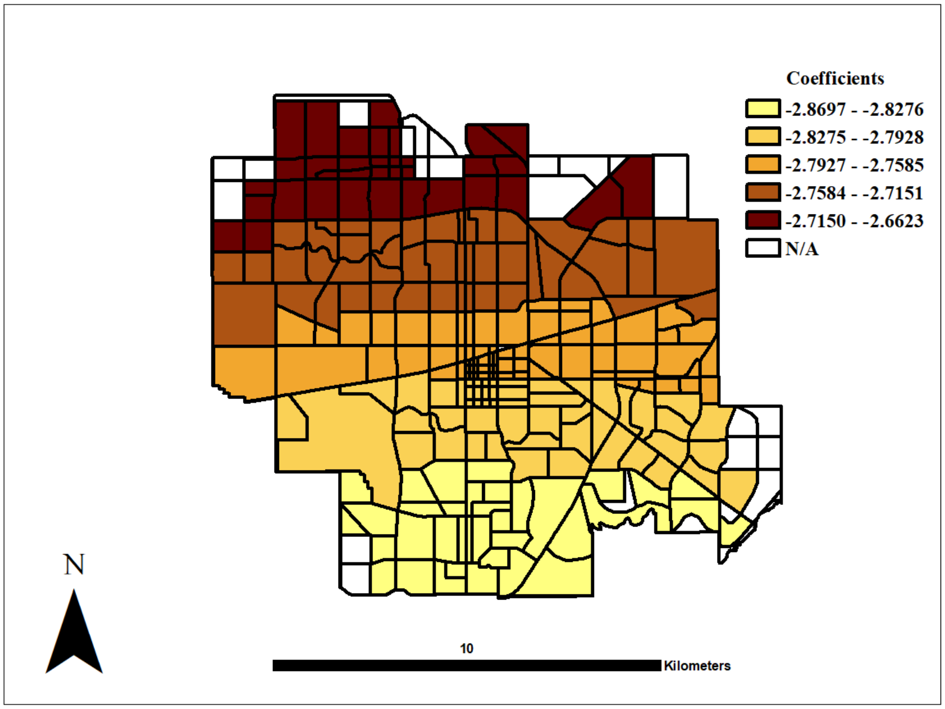


b) **Fixed Gaussian Bandwidth GWNBR**

C. POPULATION 65 PLUS

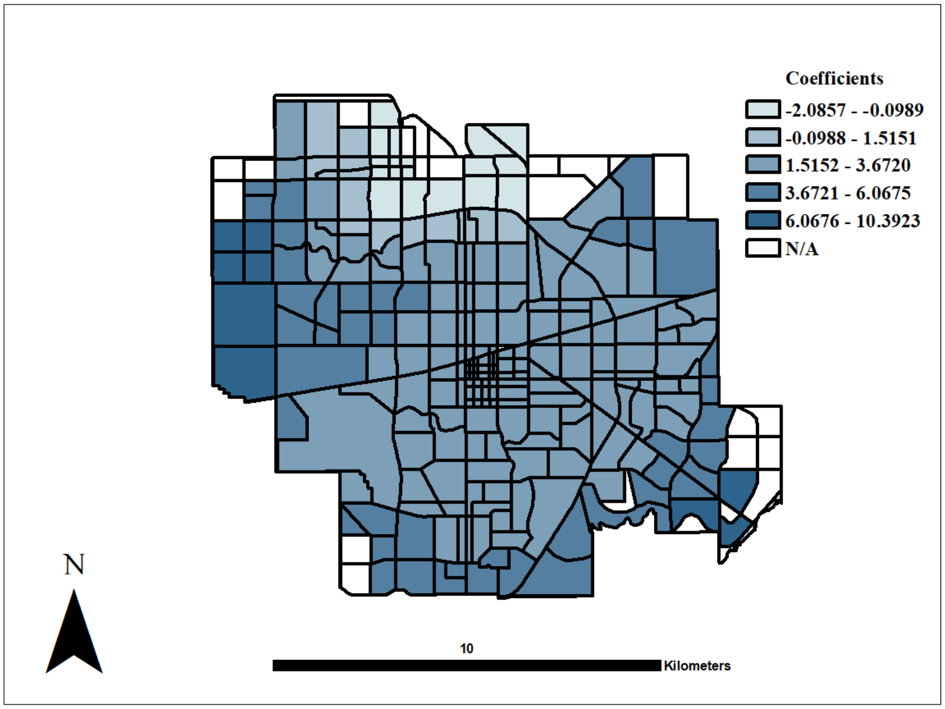


a) **Fixed Gaussian Bandwidth GWPR**

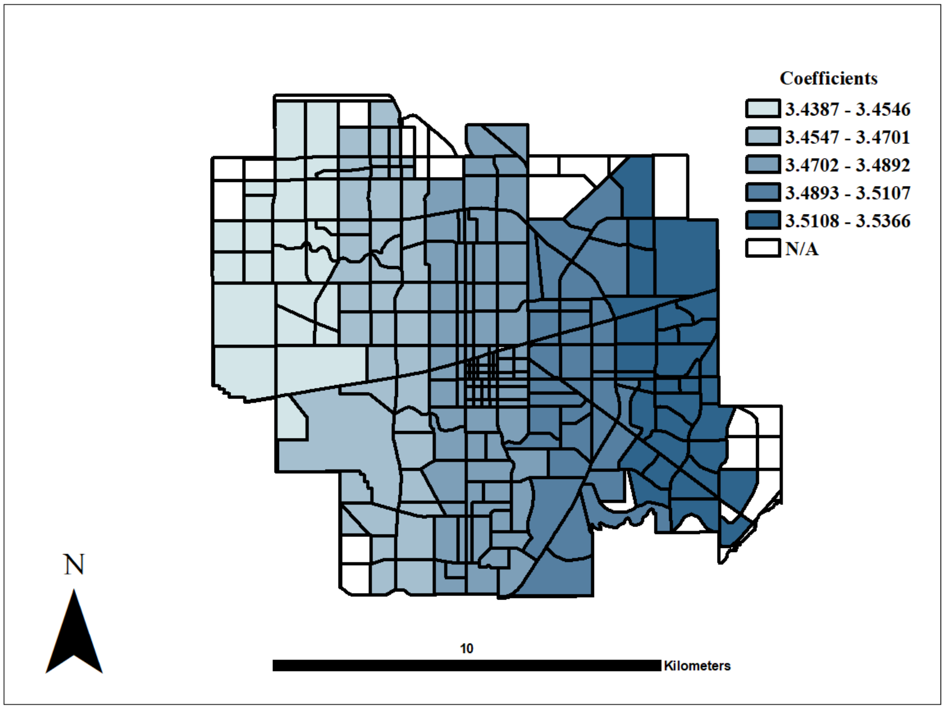


b) **Fixed Gaussian Bandwidth GWNBR**

D. RETAIL SPACE

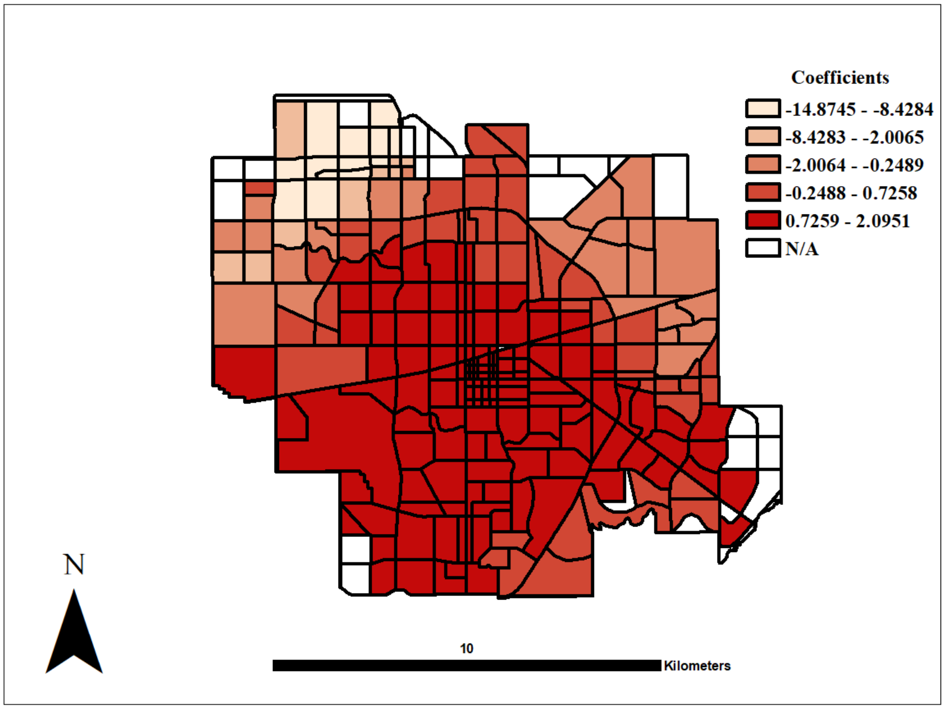


a) **Fixed Gaussian Bandwidth GWPR**

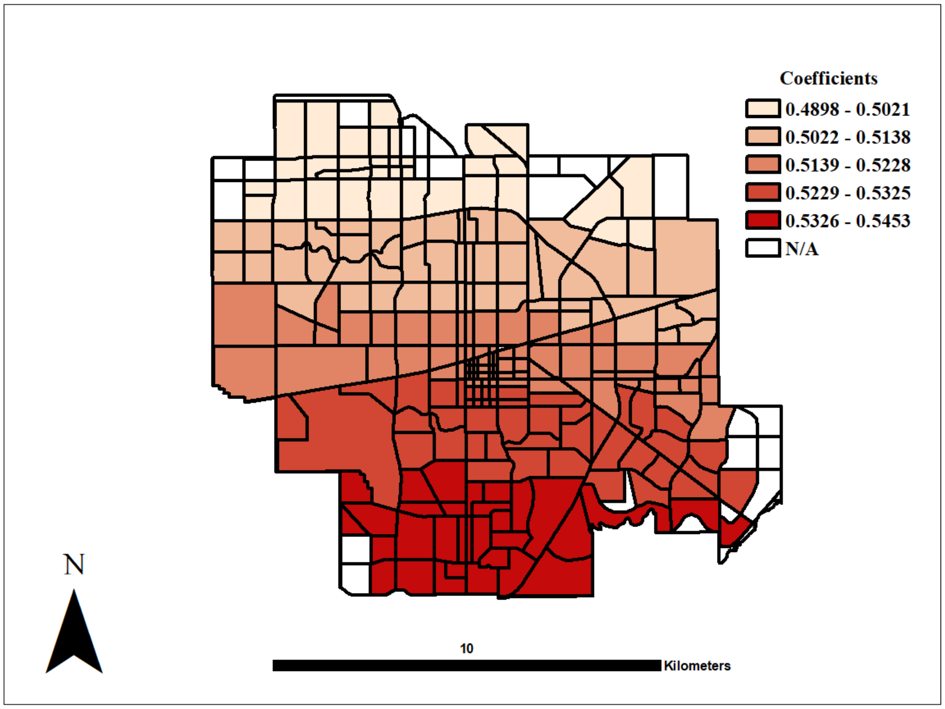


b) **Fixed Gaussian Bandwidth GWNBR**

E. OFFICE SPACE

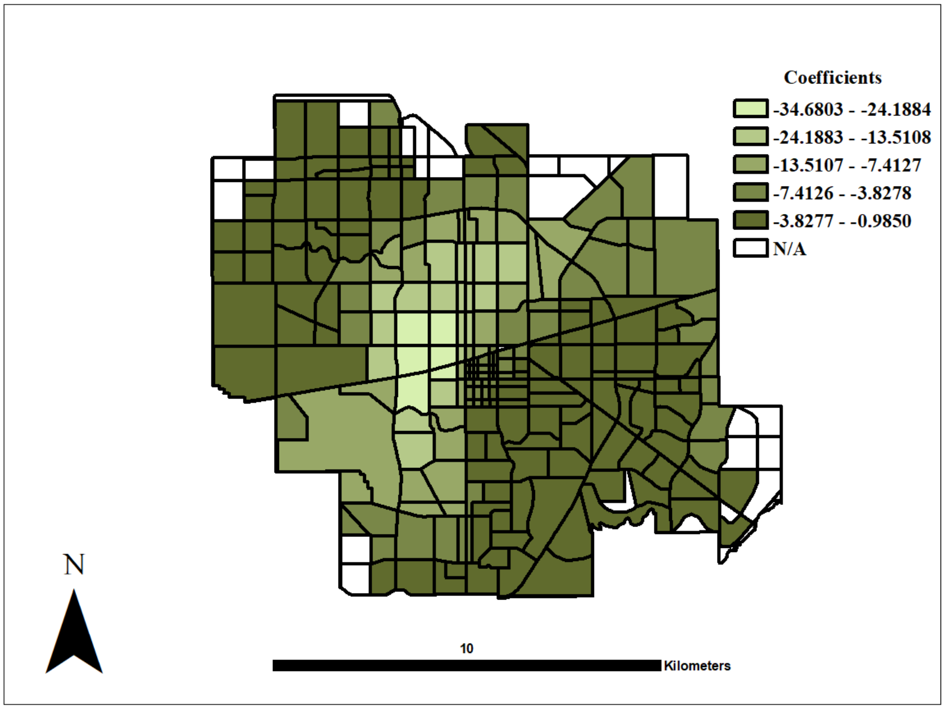


Fixed Gaussian Bandwidth GWPR

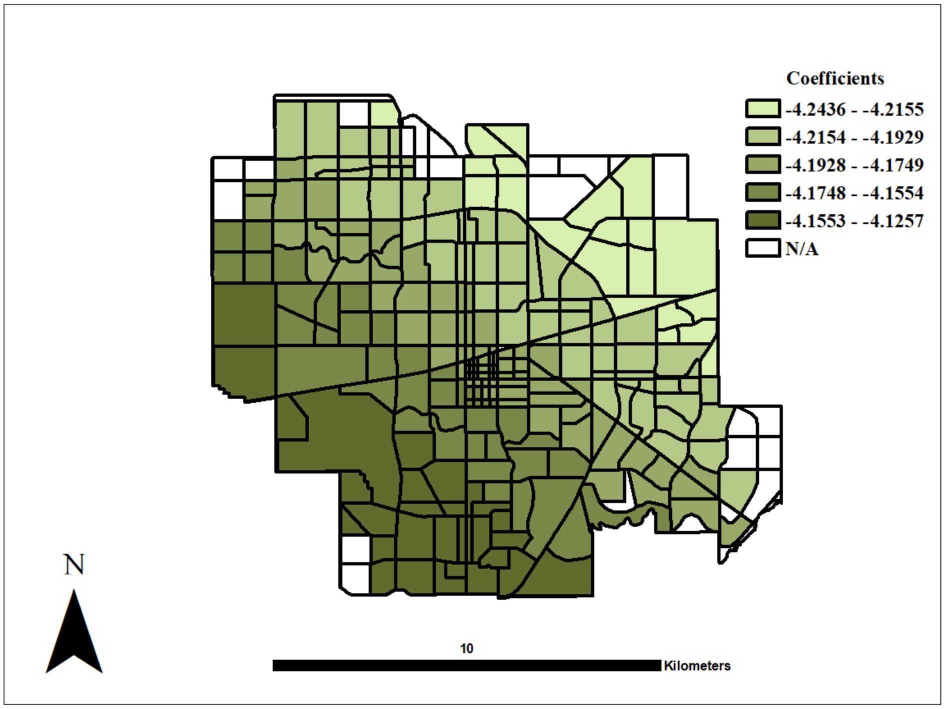


Fixed Gaussian Bandwidth GWNBR

F. URBAN HOLDING AREA

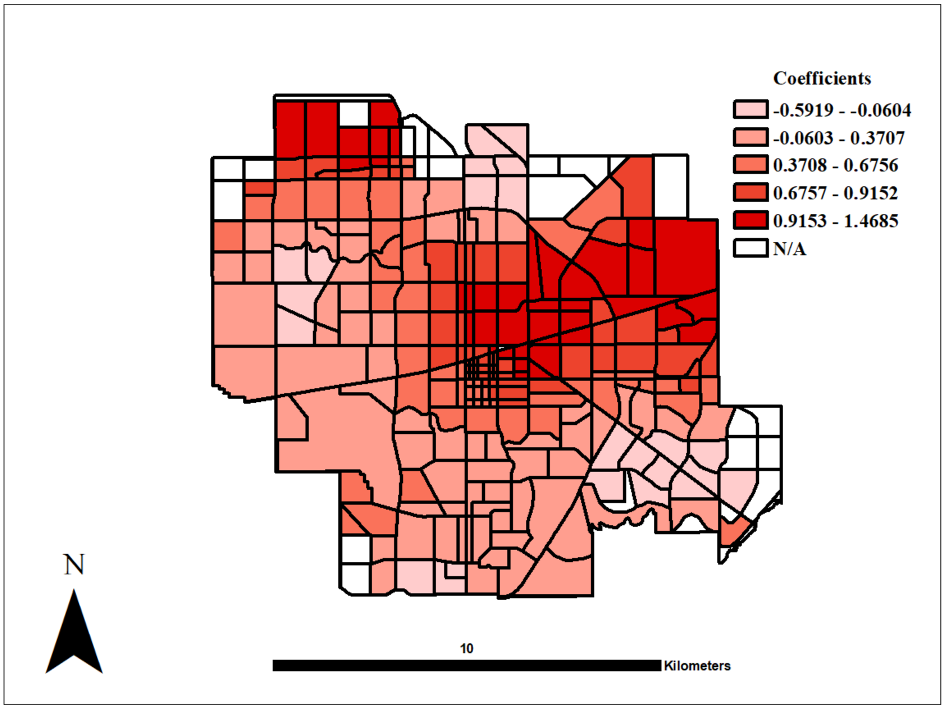


a) Fixed Gaussian Bandwidth GWPR

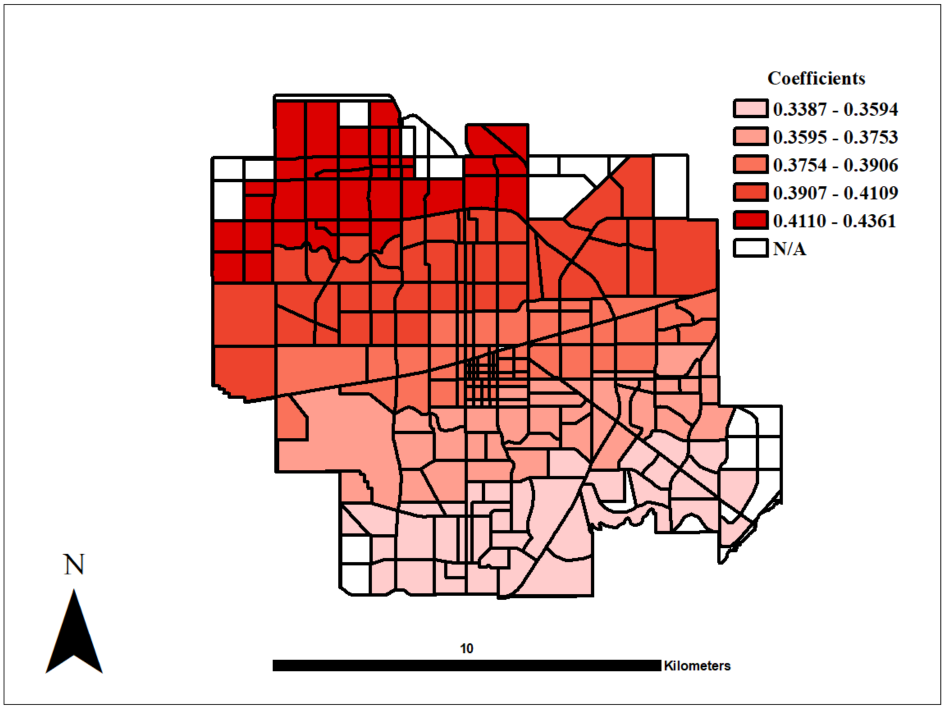


b) Fixed Gaussian Bandwidth GWNBR

G. TAZ AREA



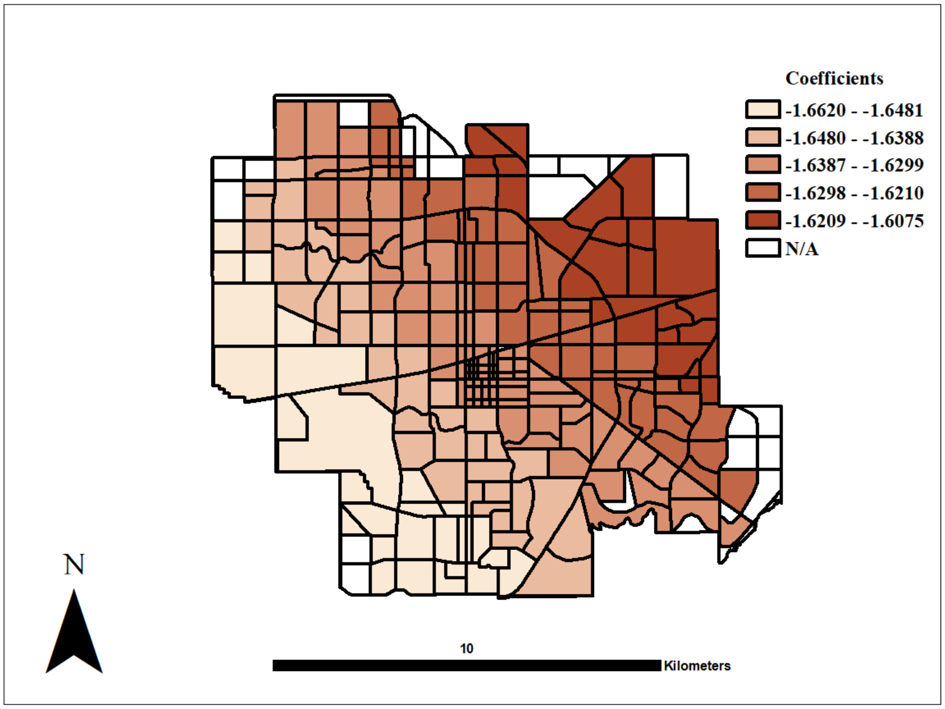
a) Fixed Gaussian Bandwidth GWPR



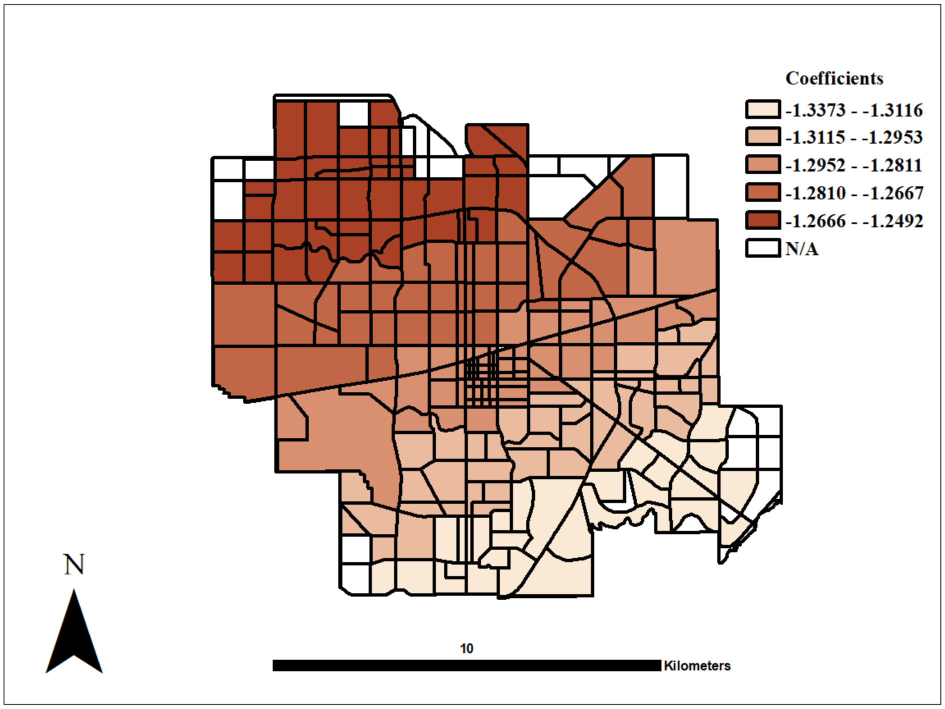
b) Fixed Gaussian Bandwidth GWNBR

Appendix I: Thematic Representation of the Coefficients of GWPR and GWNBR Models for Violent Crimes

A. INTERCEPT

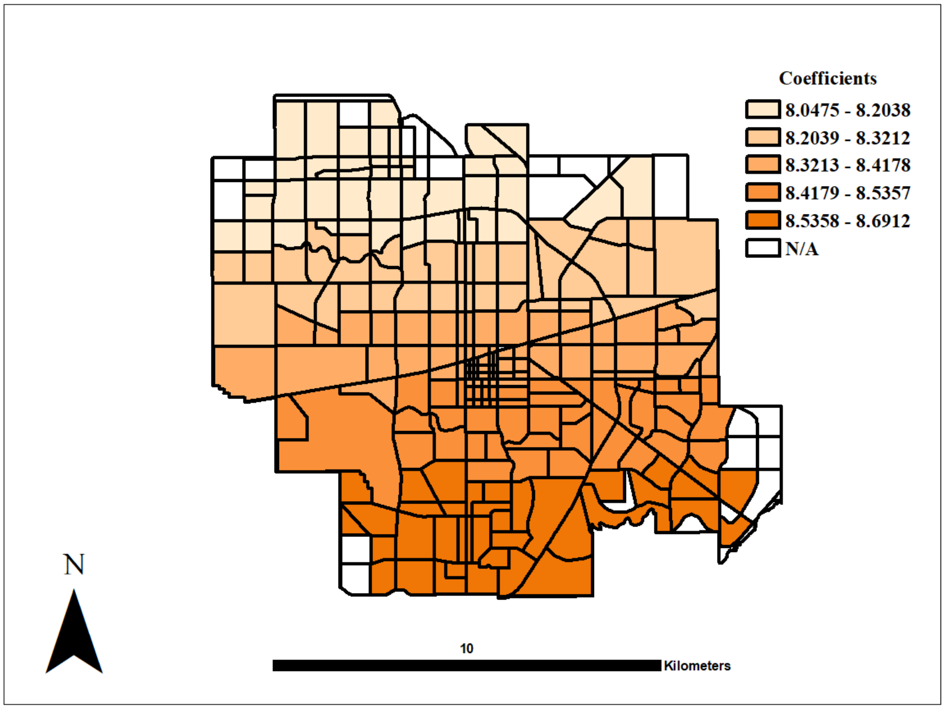


a) Fixed Gaussian Bandwidth GWPR

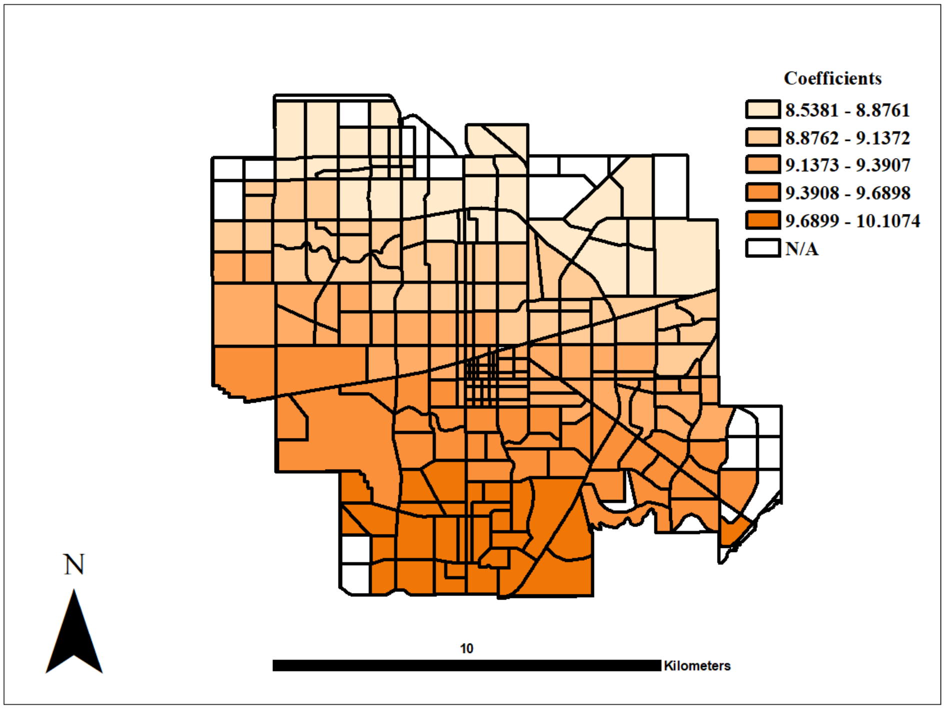


b) Fixed Gaussian Bandwidth GWNBR

B. INDUSTRY SPACE

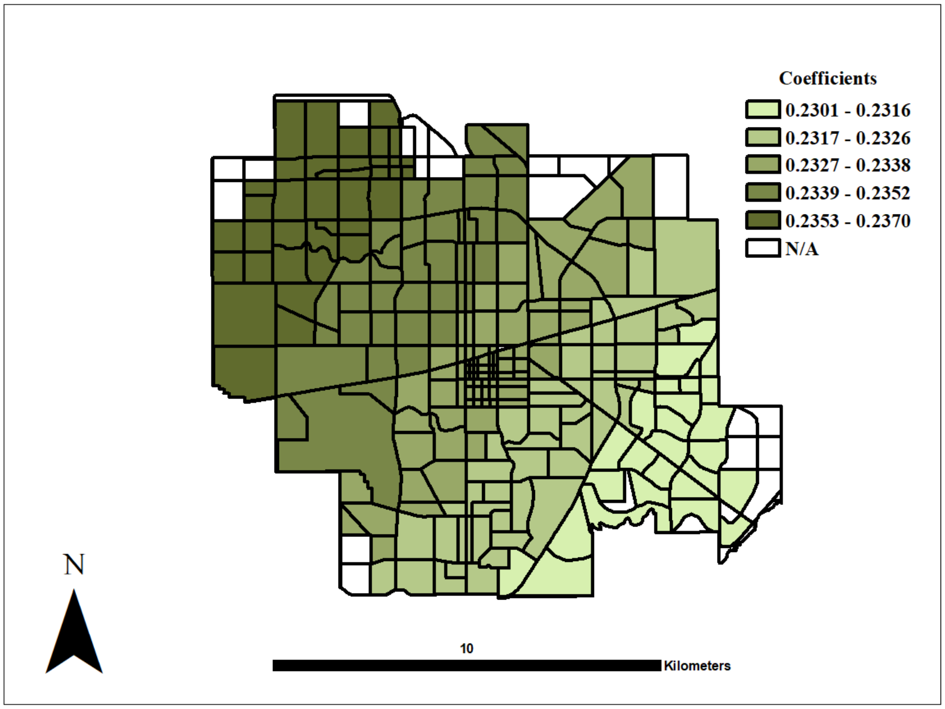


a) Fixed Gaussian Bandwidth GWPR

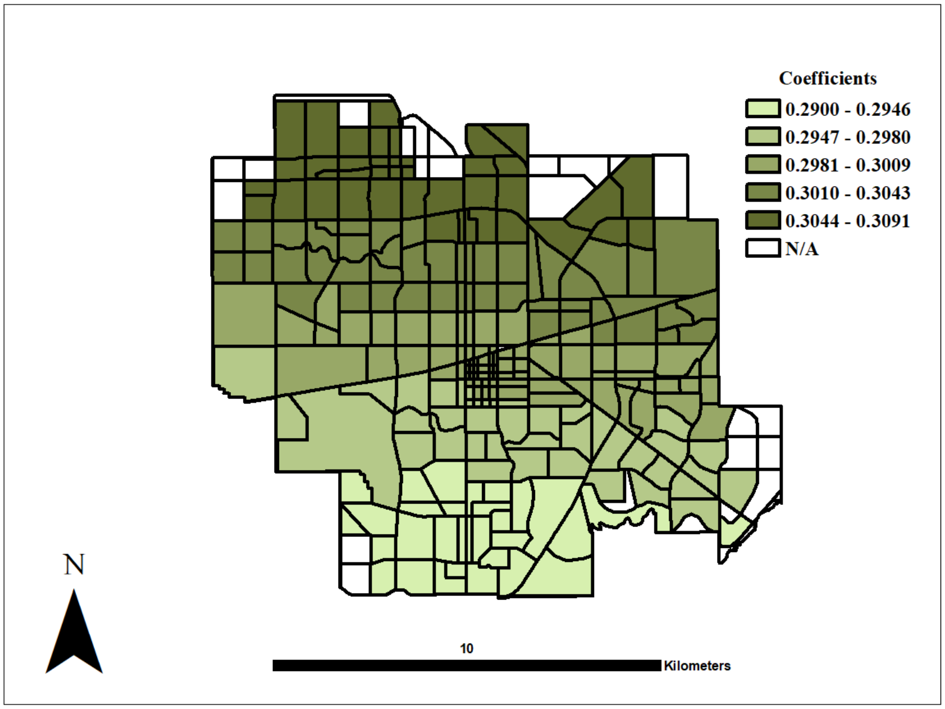


b) Fixed Gaussian Bandwidth GWNBR

C. NUMBER OF LAND USES

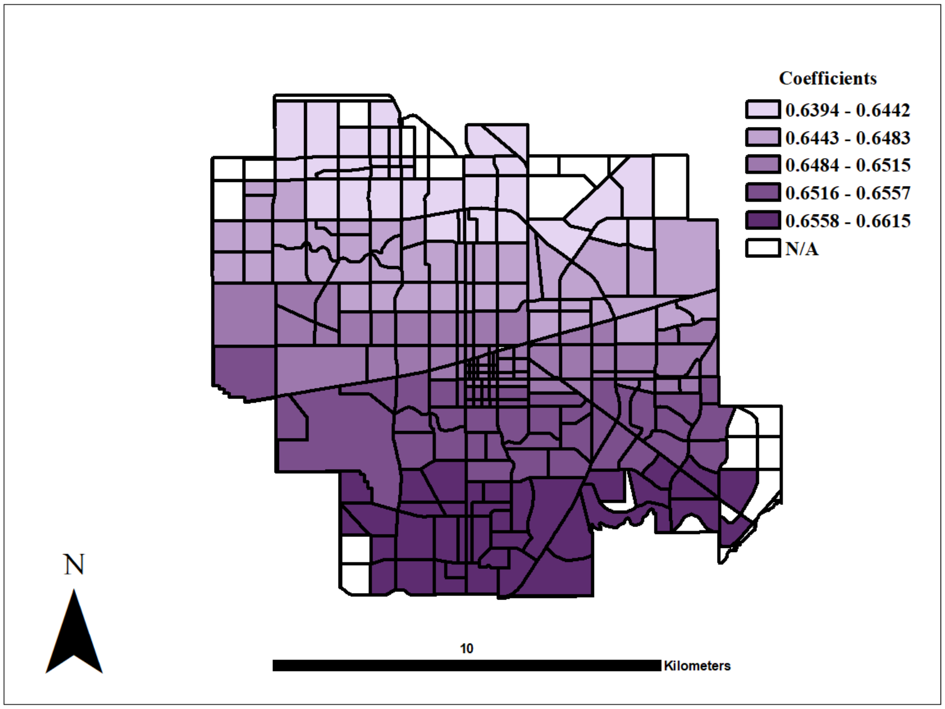


a) Fixed Gaussian Bandwidth GWPR

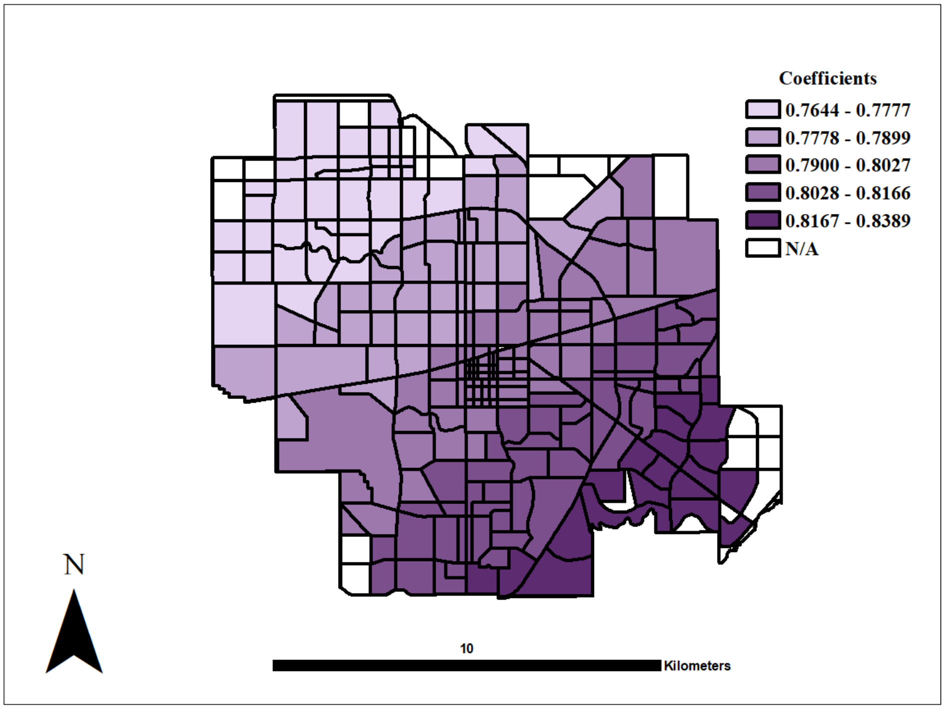


b) Fixed Gaussian Bandwidth GWNBR

D. COMMERCIAL AREA

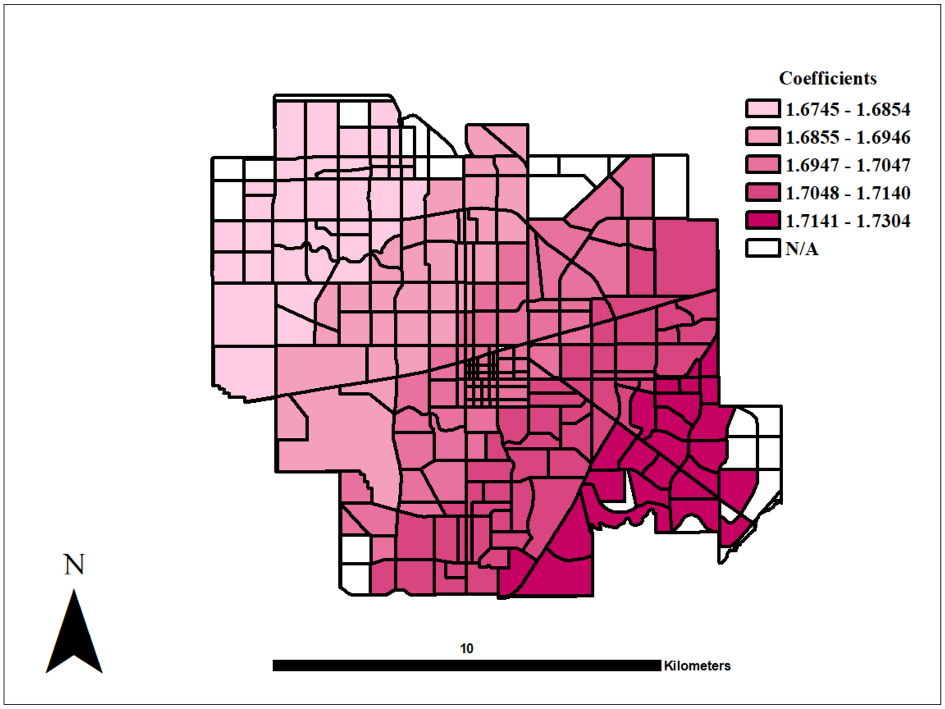


a) **Fixed Gaussian Bandwidth GWPR**

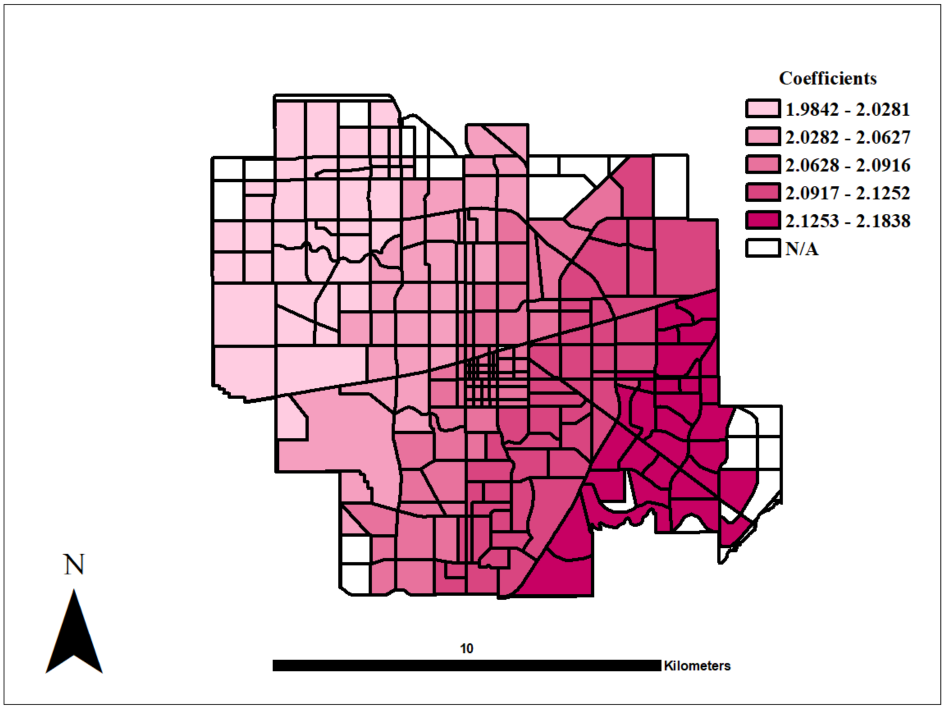


b) **Fixed Gaussian Bandwidth GWNBR**

E. RESIDENTIAL MD AREA

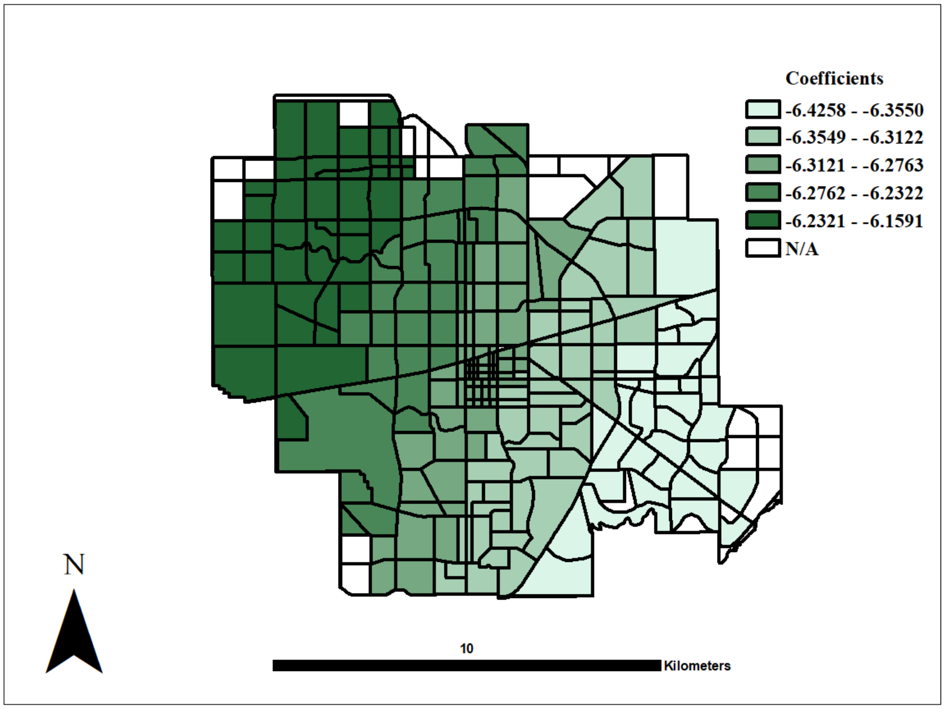


a) Fixed Gaussian Bandwith GWPR

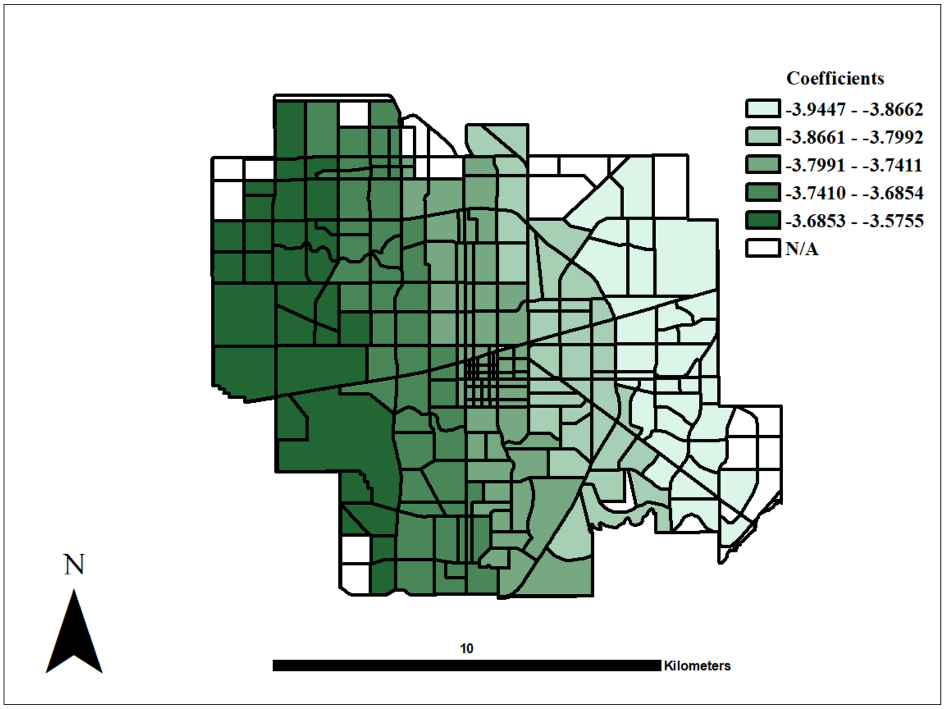


b) Fixed Gaussian Bandwidth GWNBR

F. URBAN HOLDING AREA



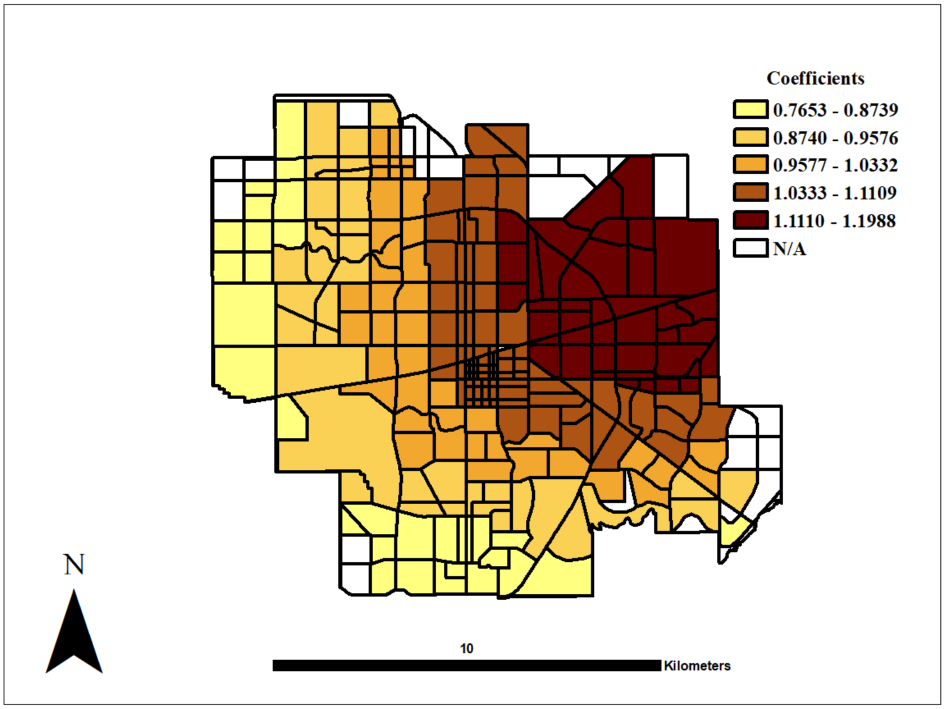
a) **Fixed Gaussian Bandwidth GWPR**



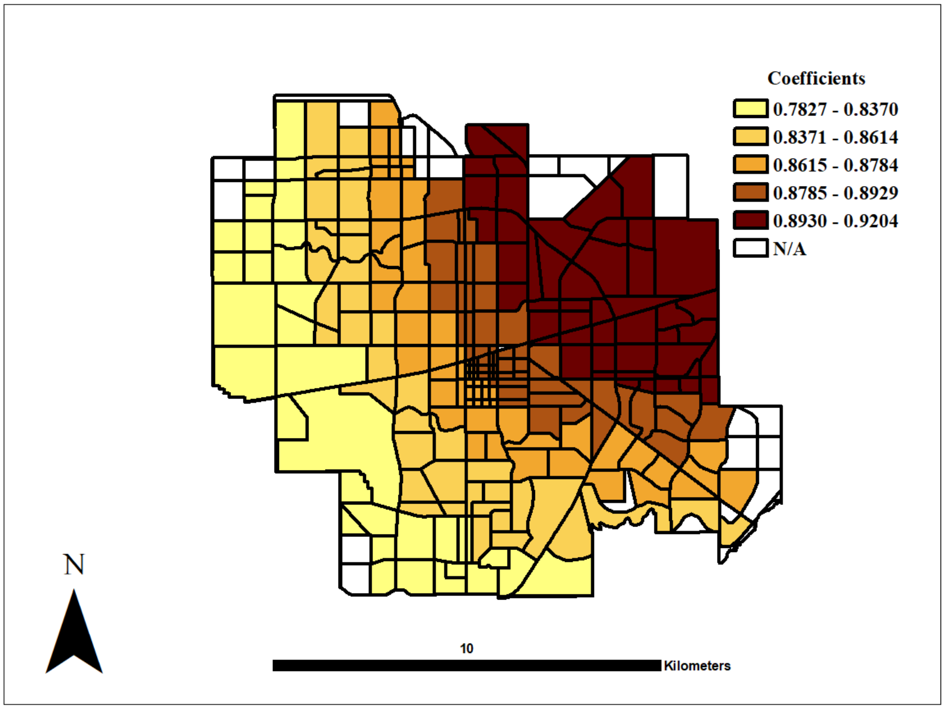
b) **Fixed Guassian Bandwidth GWNBR**

Appendix J: Thematic Representation of the Coefficients of GWPR and GWNBR Models for Non-Violent Crimes

A. INTERCEPT

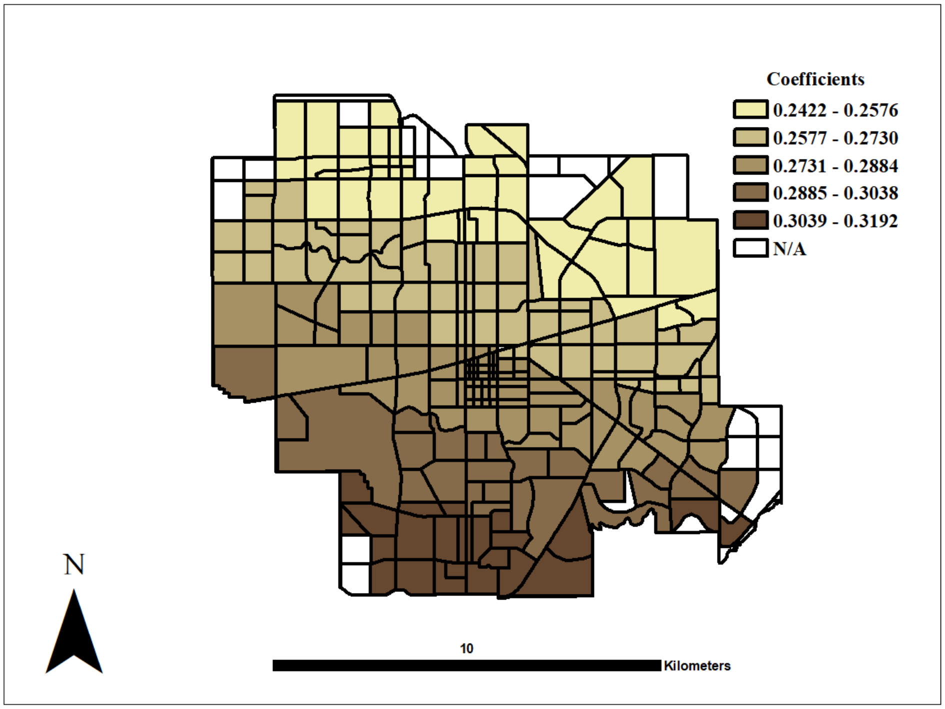


a) Fixed Gaussian Bandwidth GWPR

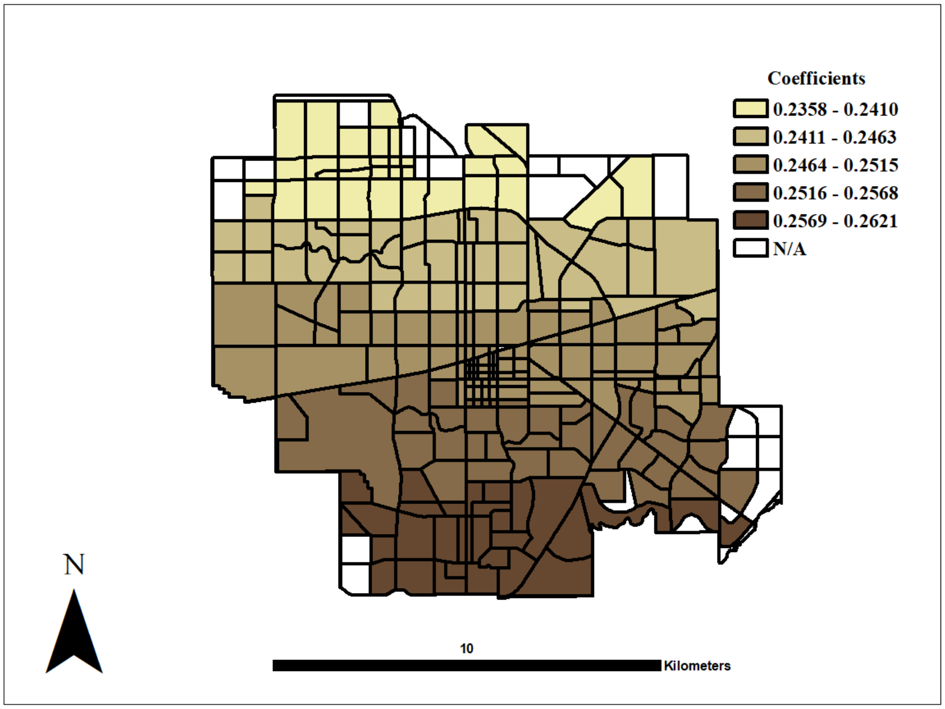


b) Fixed Gaussian Bandwidth GWNBR
c)

B. LOG TOTAL POPULATION

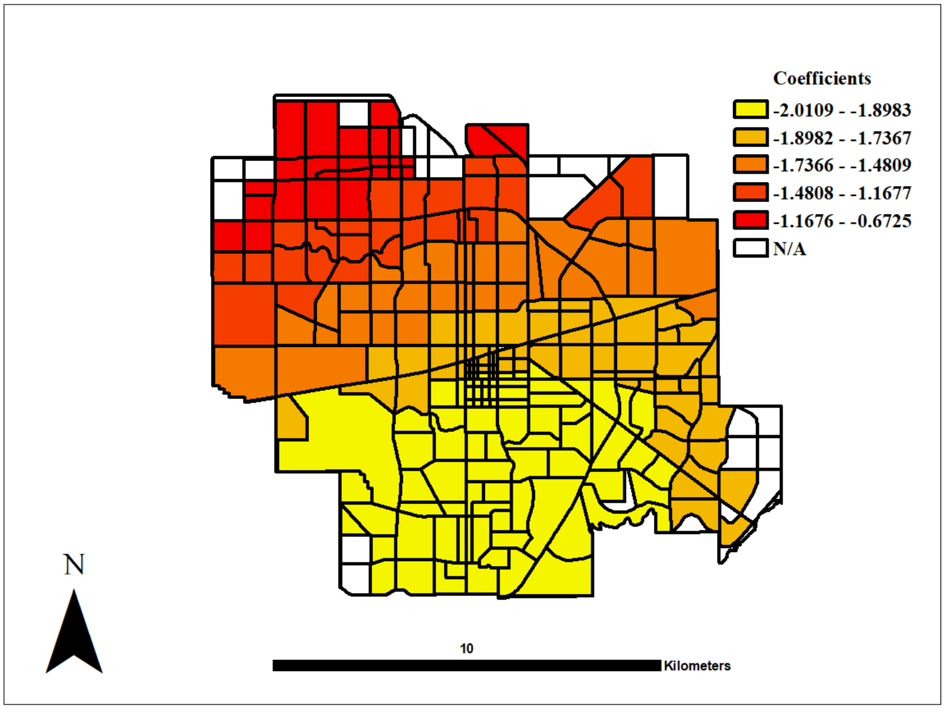


a) Fixed Gaussian Bandwidth GWPR

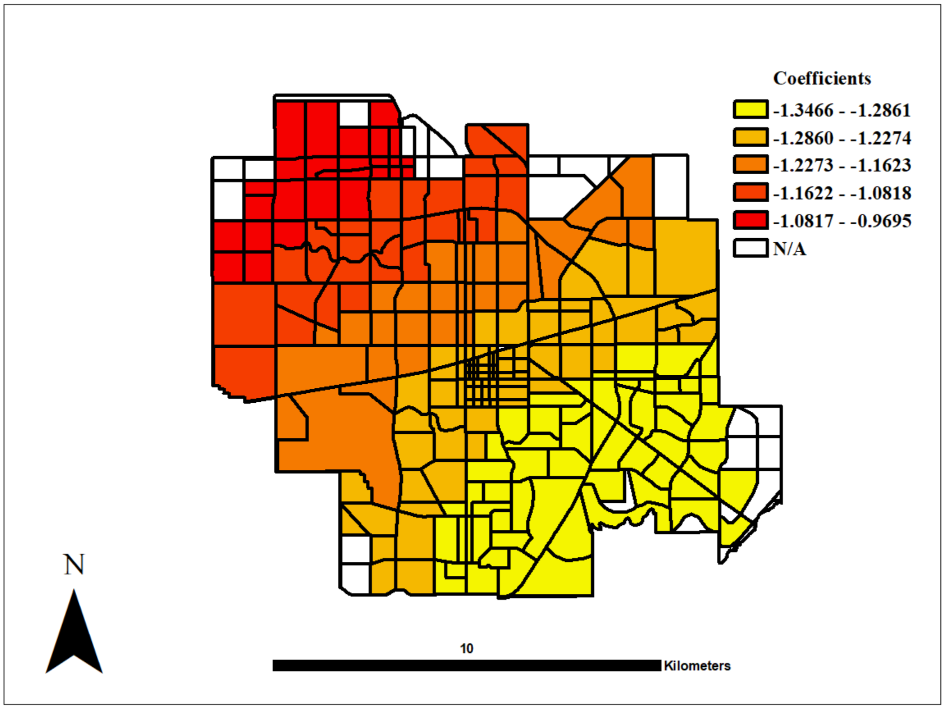


b) Fixed Gaussian Bandwidth GWNBR

C. POPULATION 65 PLUS

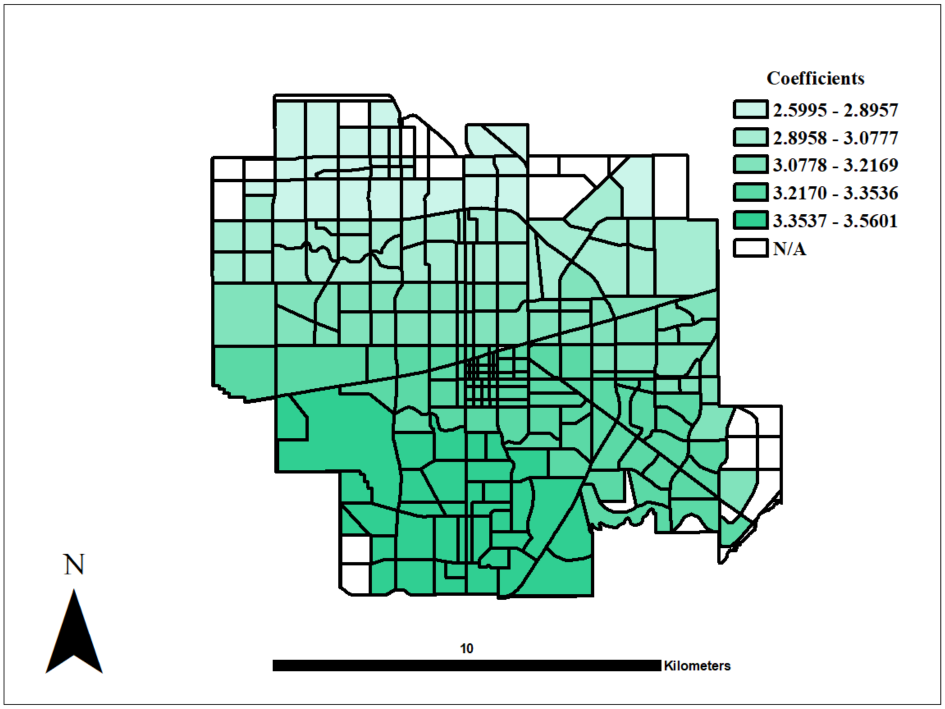


a) Fixed Gaussian Bandwidth GWPR

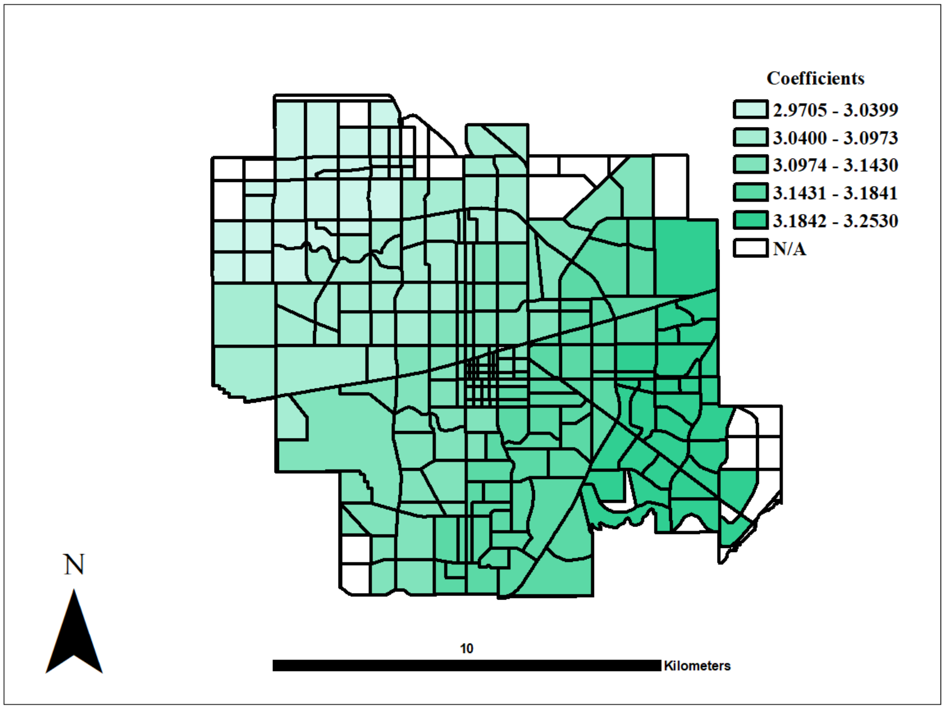


b) Fixed Gaussian Bandwidth GWNBR

D. RETAIL SPACE

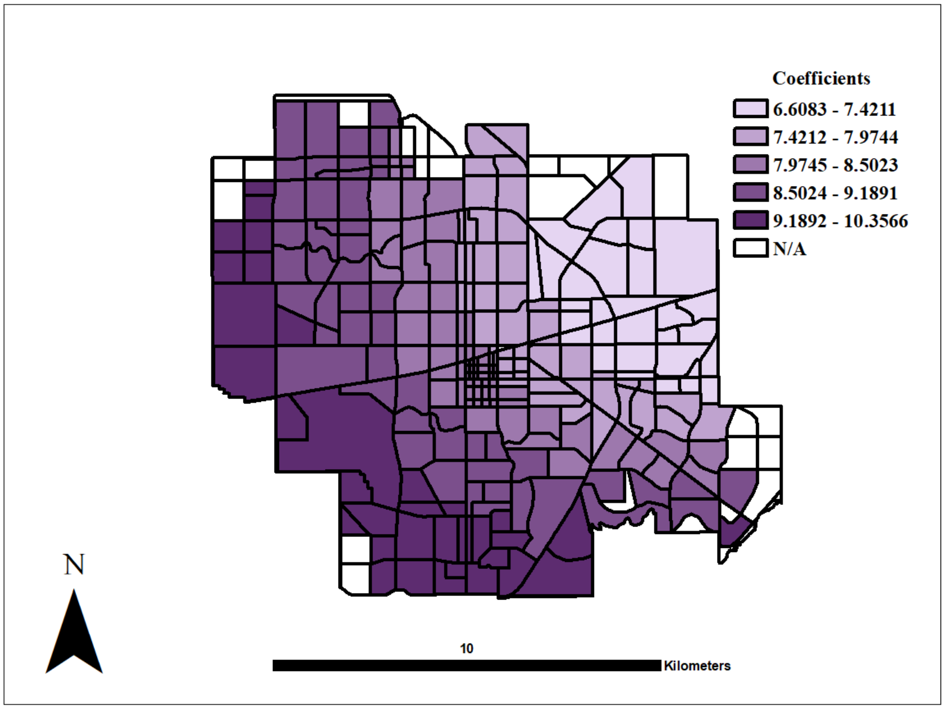


a) Fixed Gaussian Bandwidth GWPR

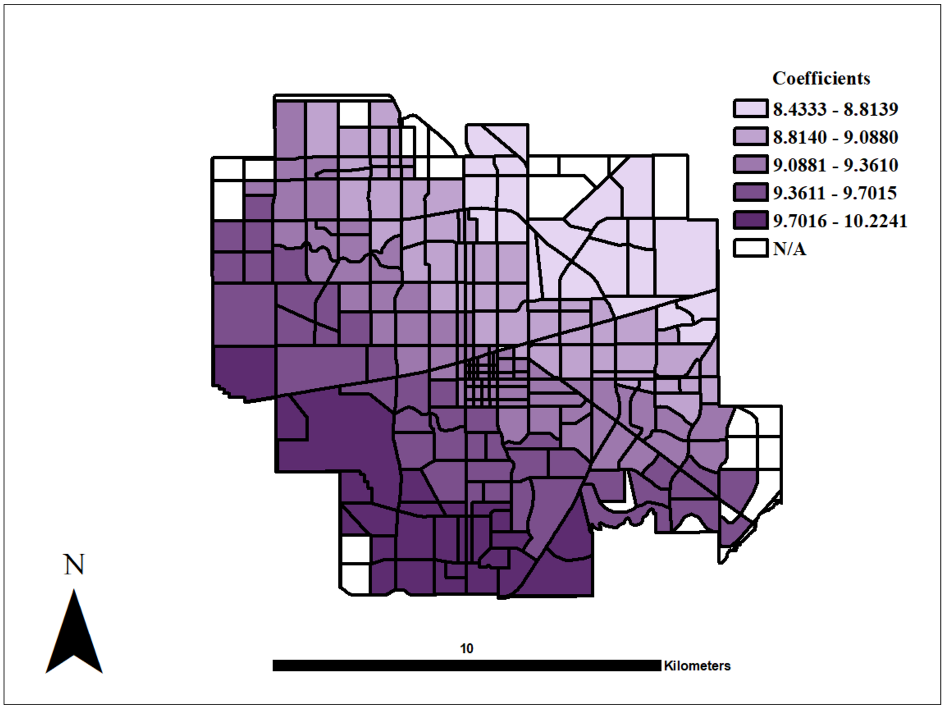


b) Fixed Gaussian Bandwidth GWNBR

E. INDUSTRY SPACE

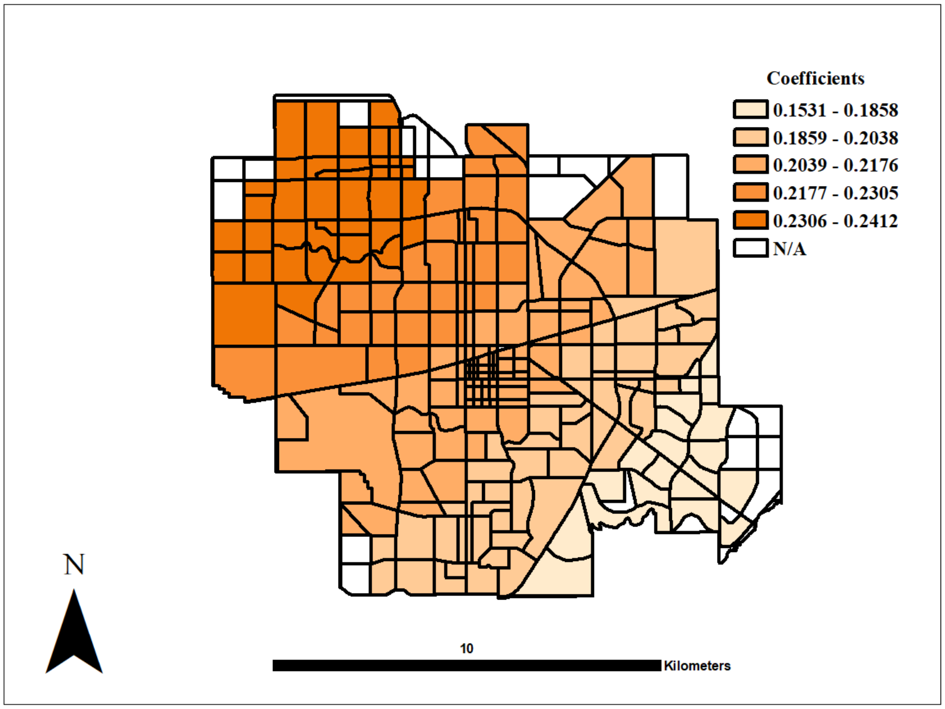


a) Fixed Gaussian Bandwidth GWPR

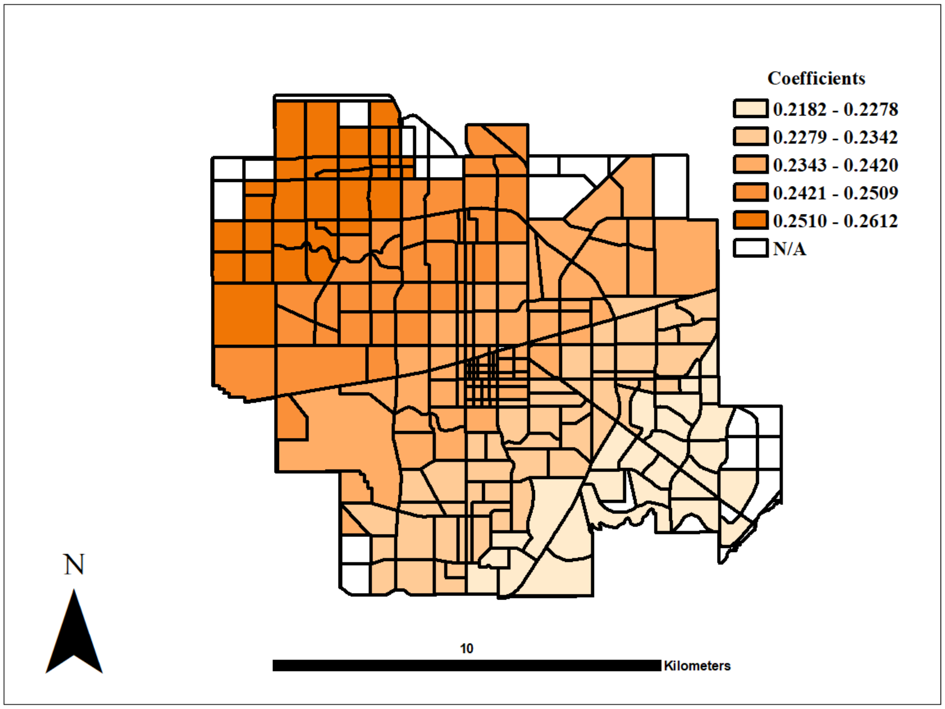


b) Fixed Gaussian Bandwidth GWNBR

F. NUMBER OF LAND USES

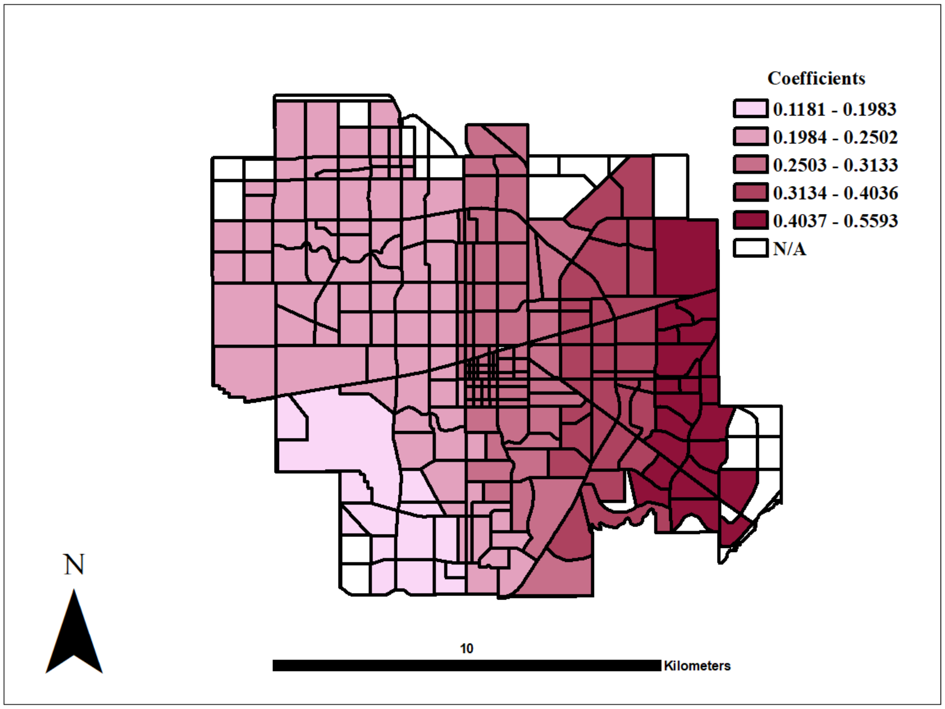


a) Fixed Gaussian Bandwidth GWPR

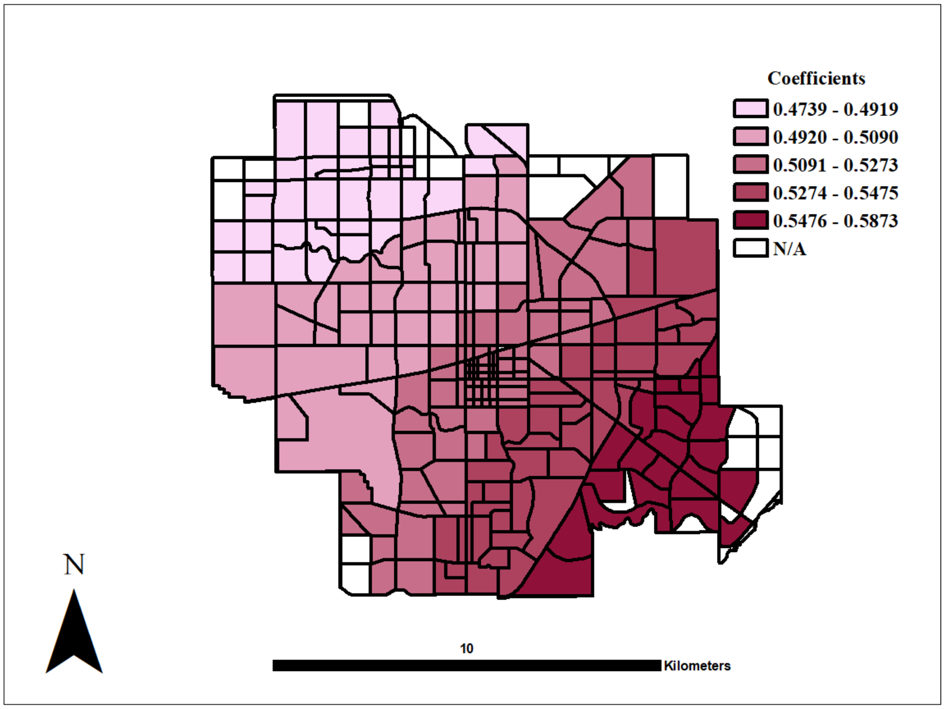


b) Fixed Gaussian Bandwidth GWNBR

G. COMMERCIAL AREA

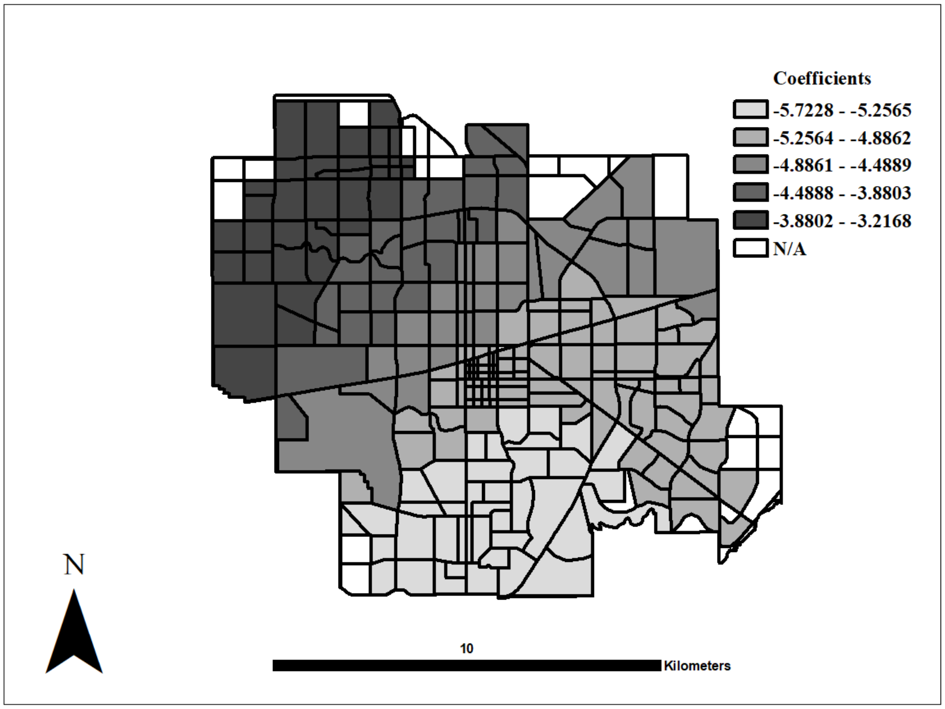


a) Fixed Gaussian Bandwidth GWPR

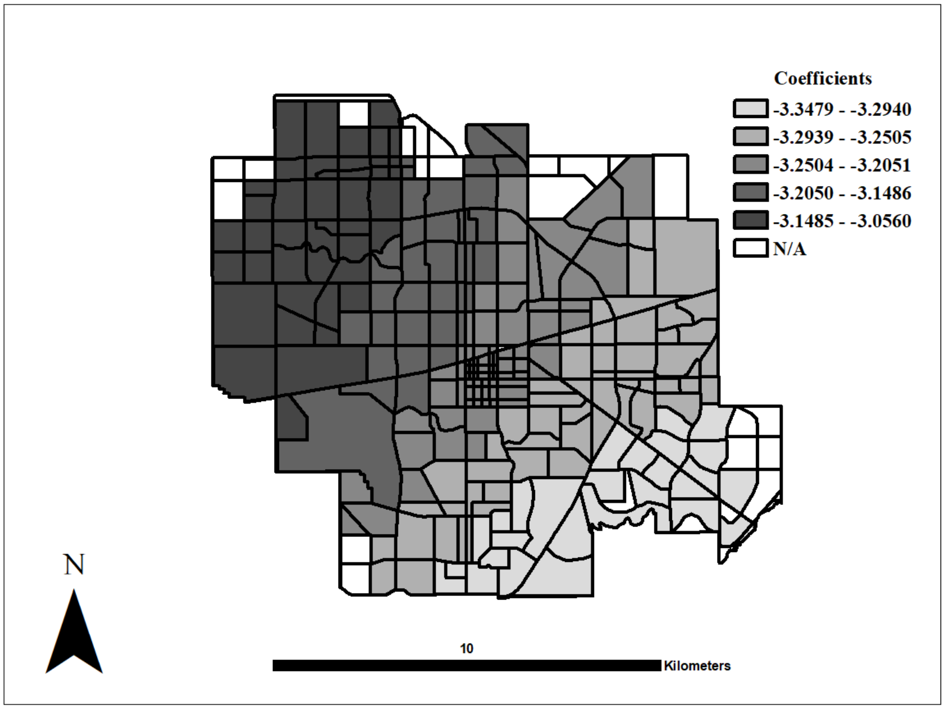


b) Fixed Gaussian Bandwidth GWNBR

H. URBAN HOLDING

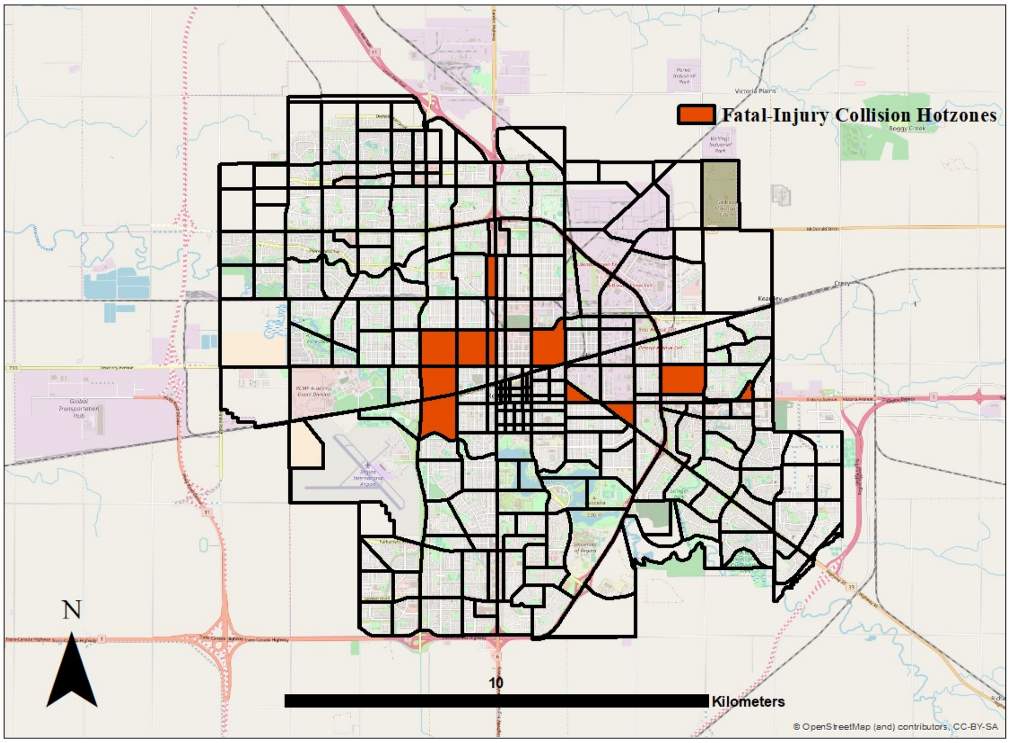


a) Fixed Gaussian Bandwidth GWPR



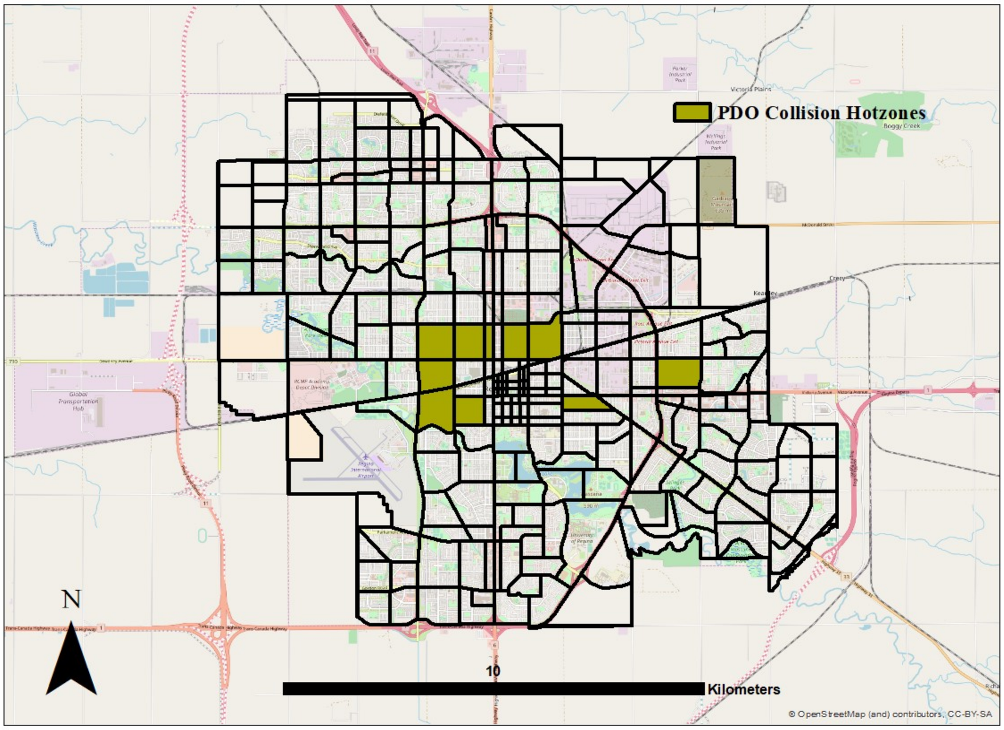
b) Fixed Gaussian Bandwidth GWNBR

Appendix K: Top Ten Location for Fatal Injury and Property Damage Only Collisions



© OpenStreetMap (and) contributors, CC-BY-SA

a) Fatal Injury Collisions Hotzone



© OpenStreetMap (and) contributors, CC-BY-SA

b) PDO Collisions Hotzone

Appendix L: Total Crimes and Total Collisions Models for DDACTS Evaluation

Table L-1: Comparison of Goodness of Fit for Total Crimes Models using GWPR and GWNBR

Crime Category	Model	AIC	AICc	BIC	MSE	MSPE	MPB	MAD
Total Crimes	**GWPR**	14691.56	14714.28	14854.94	865.96	841.12	-0.78	16.54
	GWNBR	2924.82	2925.51	2954.63	2037.70	1979.25	-0.93	25.36

Table L-2: Goodness of Fits for Total Collision Models using GWPR and GWNBR

Model	Method	Log Likelihood	AIC	AICc	BIC	MSE	MSPE	MPB	MAD
GWPR	Fixed Gaussian	-2606.64	5311.38	5336.69	5482.92	**74.29**	**71.55**	**-0.07**	**6.18**
	Adaptive Bi-square	-3167.39	6403.76	6415.50	6524.38	88.67	85.40	-0.81	7.00
GWNBR	Fixed Gaussian	**-1263.90**	**2580.40**	**2587.03**	**2672.39**	104.76	100.90	-0.28	7.33
	Adaptive Bi-square	-1271.14	2595.40	2602.16	2688.26	103.68	99.85	-0.84	7.40

Appendix M: DDACTS Zones for Enforcement Prioritization

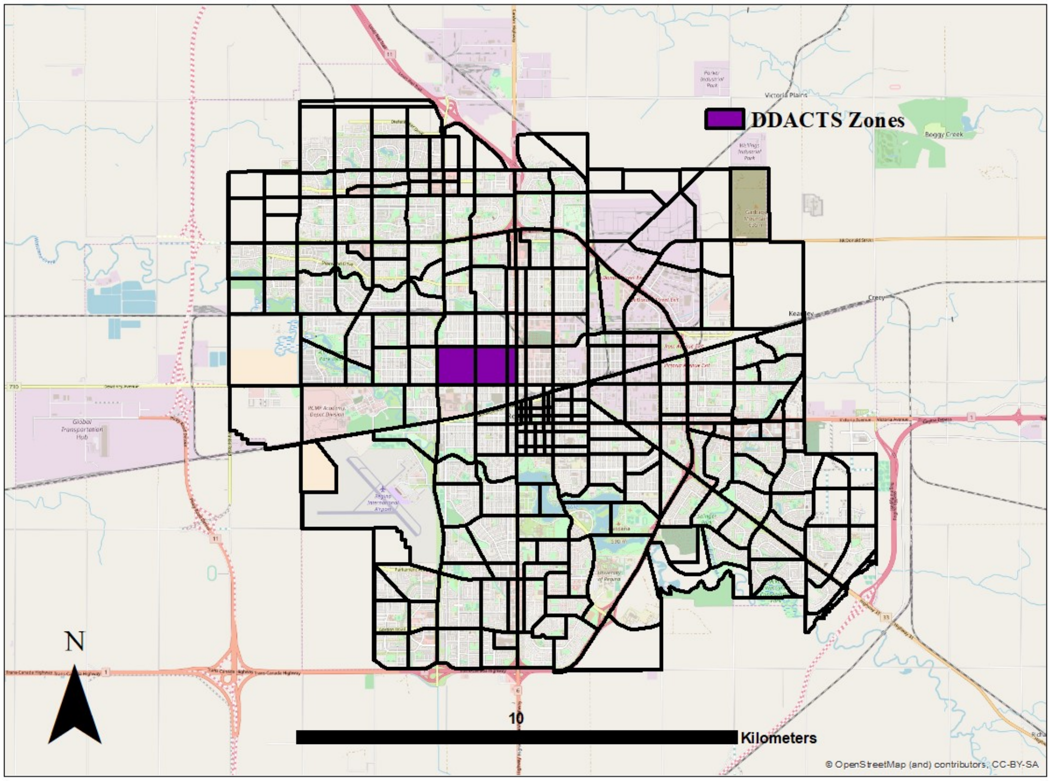


Figure M-1: Total Crimes and Fatal Injury Collisions

Table M-1: Comparison between the 244 TAZ zones, the Top 10 Hotzones and the 3 DDACTS Zones for Total Crimes and Fatal Injury Collisions

Categories	Total Crimes	Fatal Injury Collisions
All Zones numbers of incidences (Crimes and Collisions)	50284 crimes	5759 collisions
Top 10 Zones	10633 crimes	965 collisions
DDACTS hotzones	4779 crimes	261 collisions
Top 10 hotzones/All 244 Zones	21.15%	16.76%
DDACTS zones/All 244 zones	9.50%	4.53%
Area of all zones	147.93 km^2	147.93 km^2
Area of Top 10 hotzones (km^2)	4.04 km^2	5.34 km^2
Area of 3 DDACTS zones	1.57 km^2	1.57 km^2
Area of Top 10 hotzones/area of all zones	2.73%	3.61%
Area of 3 DDACTS hotzones/area of all zones	1.1%	1.1%

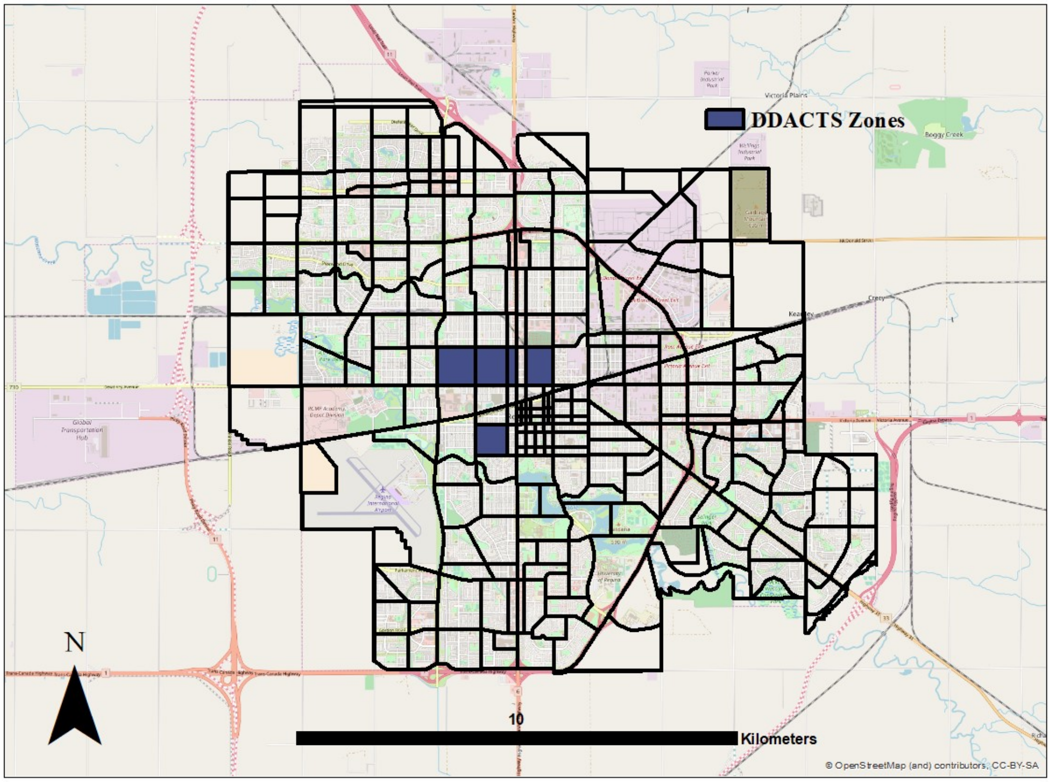


Figure M-2: Total Crimes and Property Damage Only Collisions

Table M-2: Comparison between the 244 TAZ zones, the Top 10 Hotzones and the 5 DDACTS Zones for Total Crimes and Property Damage Only Collisions

Categories	Total Crimes	Property Damage Only Collisions
All Zones numbers of incidences (Crimes and Collisions)	50284 crimes	20883 collisions
Top 10 Zones	10633 crimes	2769 collisions
5 DDACTS hotzones	6245 crimes	1459 collisions
Top 10 hotzones/All 244 Zones	21.1%	13.3%
DDACTS zones/All 244 zones	12.42%	6.99%
Area of all zones	147.93 km^2	147.93 km^2
Area of Top 10 hotzones (km^2)	4.04 km^2	5.92 km^2
Area of 5 DDACTS zones	2.61 km^2	2.61 km^2
Area of Top 10 hotzones/area of all zones	2.73%	4.00%
Area of 5 DDACTS hotzones/area of all zones	1.76%	1.76%

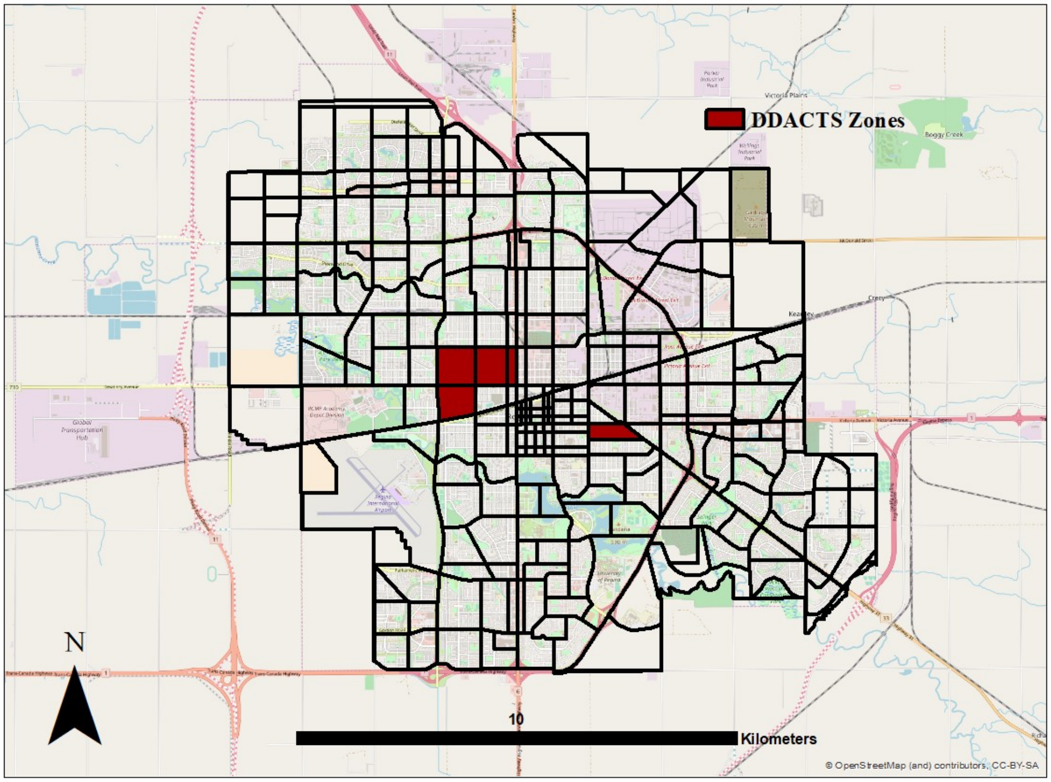


Figure M-3: Violent Crimes and Total Collisions

Table M-3: Comparison between the 244 TAZ zones, the Top 10 Hotzones and the 4 DDACTS Zones for Violent Crimes and Total Collisions

Categories	Violent Crimes	Total Collisions
All Zones numbers of incidences (Crimes and Collisions)	9181 crimes	26642 collisions
Top 10 Zones	3282 crimes	3546 collisions
4 DDACTS hotzones	2045 crimes	1829 collisions
Top 10 hotzones/All 244 Zones	35.75%	13.31%
DDACTS zones/All 244 zones	22.27%	6.87%
Area of all zones	147.93 km^2	147.93 km^2
Area of Top 10 hotzones (km^2)	4.49 km^2	5.58 km^2
Area of 4 DDACTS zones	2.56 km^2	2.56 km^2
Area of Top 10 hotzones/area of all zones	3.04%	3.77%
Area of 4 DDACTS hotzones/area of all zones	1.73%	1.73%

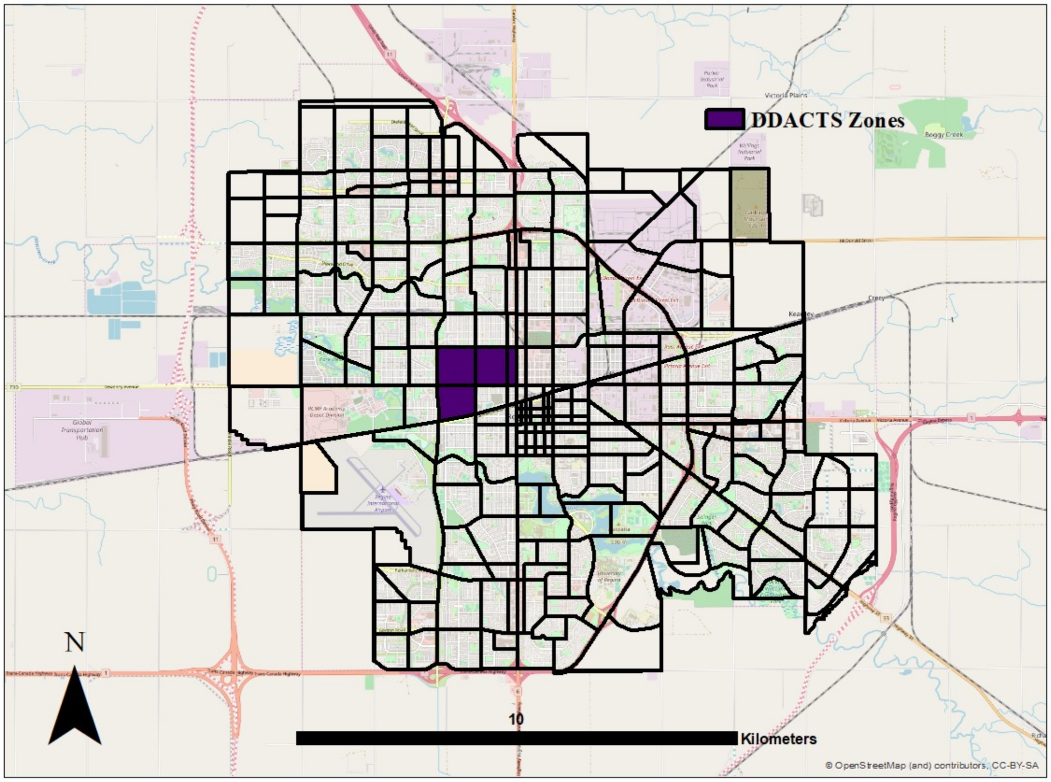


Figure M-4: Violent Crimes and Fatal Injury Collisions

Table M-4: Comparison between the 244 TAZ zones, the Top 10 Hotzones and the 4 DDACTS Zones for Violent Crimes and Fatal Injury Collisions

Categories	Violent Crimes	Fatal Injury Collisions
All Zones numbers of incidences (Crimes and Collisions)	9181 crimes	5759 collisions
Top 10 Zones	3282 crimes	965 collisions
4 DDACTS hotzones	1908 crimes	352 collisions
Top 10 hotzones/All 244 Zones	35.75%	16.76%
DDACTS zones/All 244 zones	20.78%	6.11%
Area of all zones	147.93 km^2	147.93 km^2
Area of Top 10 hotzones)	4.49 km^2	5.34 km^2
Area of 4 DDACTS zones	2.19 km^2	2.19 km^2
Area of Top 10 hotzones/area of all zones	3.04%	3.61%
Area of 4 DDACTS hotzones/area of all zones	1.48%	1.48%

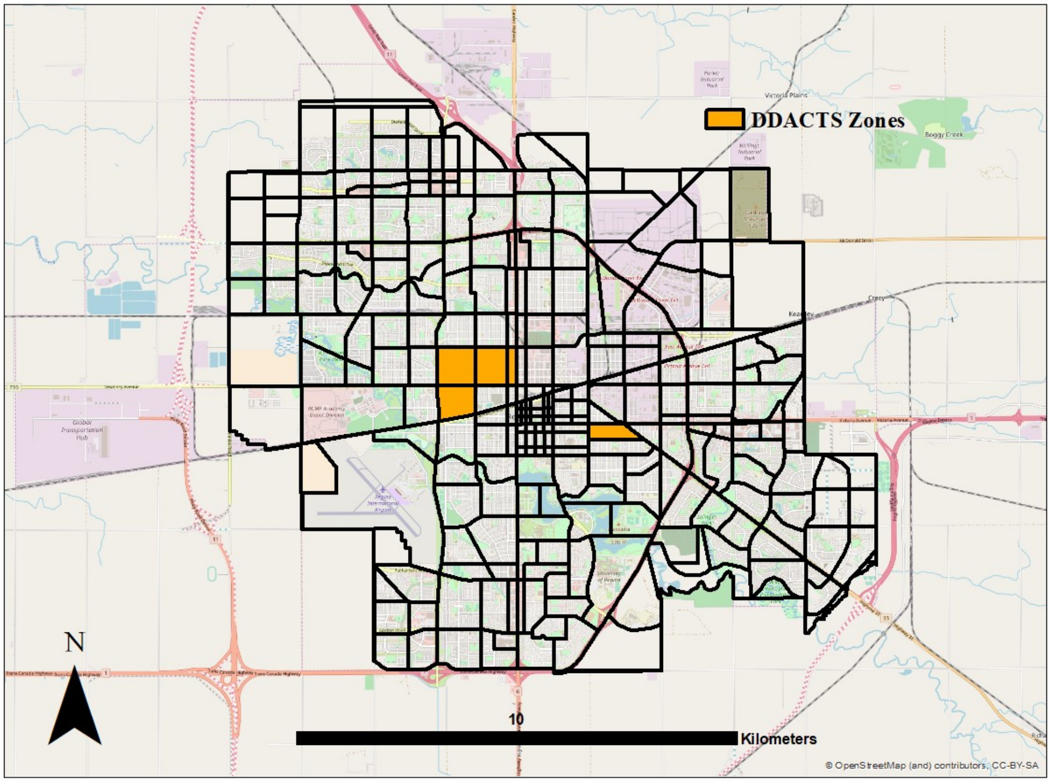


Figure M-5: Violent Crimes and Property Damage Only Collisions

Table M-5: Comparison between the 244 TAZ zones, the Top 10 Hotzones and the 5 DDACTS Zones for Violent Crimes and Property Damage Only Collisions

Categories	Violent Crimes	Property Damage Only Collisions
All Zones numbers of incidences (Crimes and Collisions)	9181 crimes	20883 collisions
Top 10 Zones	3282 crimes	2769 collisions
DDACTS hotzones	2045 crimes	1417 collisions
Top 10 hotzones/All 244 Zones	35.75%	13.26%
5 DDACTS zones/All 244 zones	22.27%	6.79%
Area of all zones	147.93 km^2	147.93 km^2
Area of Top 10 hotzones (km^2)	4.49 km^2	5.92 km^2
Area of 5 DDACTS zones	2.56 km^2	2.56 km^2
Area of Top 10 hotzones/area of all zones	3.04%	4.00%
Area of 5 DDACTS hotzones/area of all zones	1.73%	1.73%

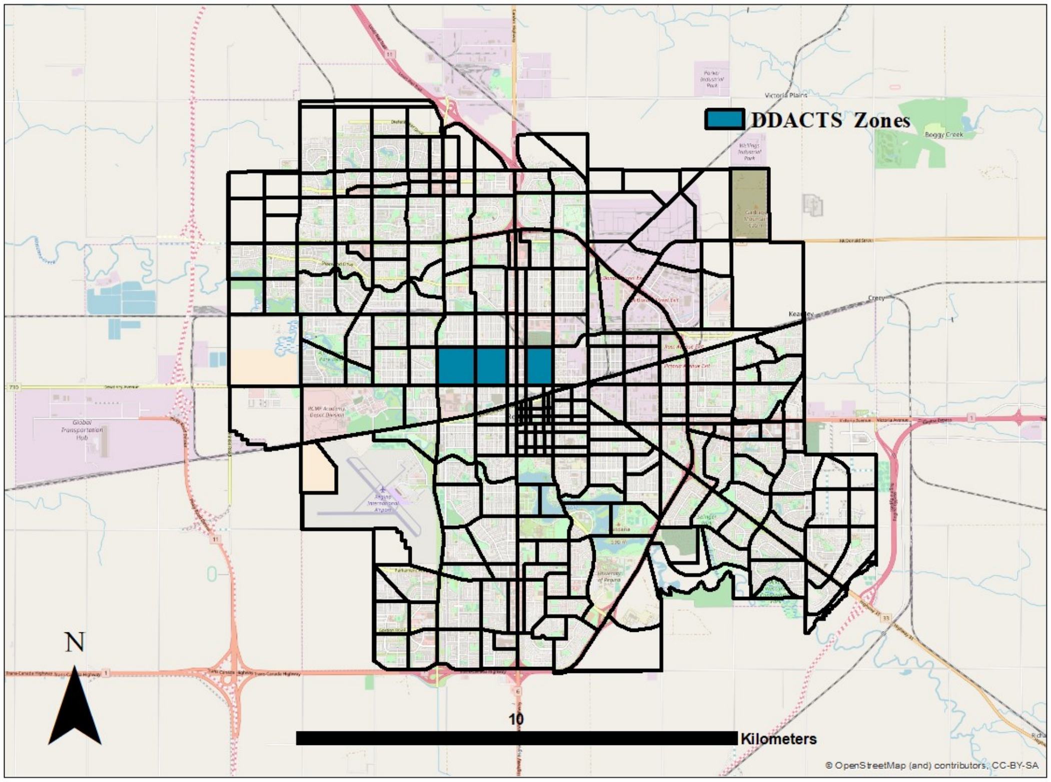


Figure M-6: Non Violent Crimes and Total Collisions

Table M-6: Comparison between the 244 TAZ zones, the Top 10 Hotzones and the 3 DDACTS Zones for Non-Violent Crimes and Total Collisions

Categories	Non-Violent Crimes	Total Collision
All Zones numbers of incidences (Crimes and Collisions)	41103 crimes	26642 collisions
Top 10 Zones	8315 crimes	3546 collisions
3 DDACTS hotzones	3195 crimes	1181 collisions
Top 10 hotzones/All 244 Zones	20.23%	13.31%
DDACTS zones/All 244 zones	7.77%	4.43%
Area of all zones	147.93 km^2	147.93 km^2
Area of Top 10 hotzones (km^2)	5.56 km^2	5.58 km^2
Area of 3 DDACTS zones	1.92 km^2	1.92 km^2
Area of Top 10 hotzones/area of all zones	3.76%	3.77%
Area of 3 DDACTS hotzones/area of all zones	1.30%	1.30%

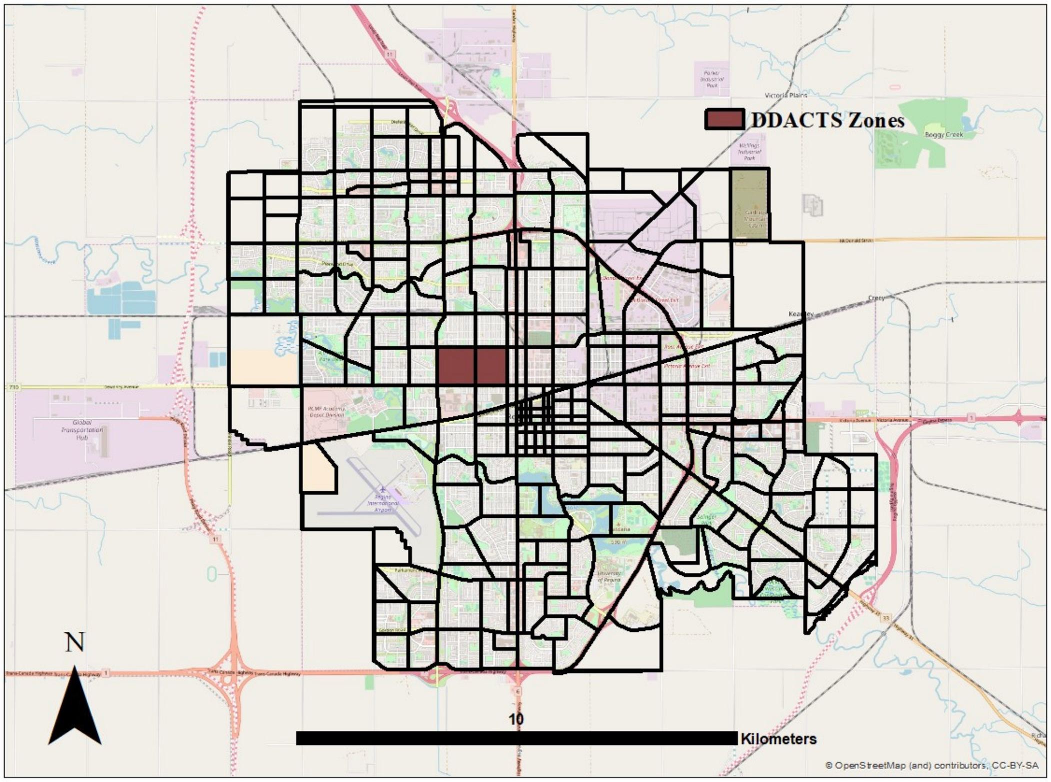


Figure M-7: Non-violent Crimes and Fatal Injury Collisions

Table M-7: Comparison between the 244 TAZ zones, the Top 10 Hotzones and the 2 DDACTS Zones for Non-Violent Crimes and Fatal Injury Collisions

Categories	Non-Violent Crimes	Fatal Injury Collisions
All Zones numbers of incidences (Crimes and Collisions)	41103 crimes	5759 collisions
Top 10 Zones	8315 crimes	965 collisions
2 DDACTS hotzones	2600 crimes	180 collisions
Top 10 hotzones/All 244 Zones	20.23%	16.76%
DDACTS zones/All 244 zones	6.33%	3.13%
Area of all zones	147.93 km^2	147.93 km^2
Area of Top 10 hotzones (km^2)	5.56 km^2	5.34 km^2
Area of 2 DDACTS zones	1.38 km^2	1.38 km^2
Area of Top 10 hotzones/area of all zones	3.76%	3.61%
Area of 2 DDACTS hotzones/area of all zones	0.94%	0.94%

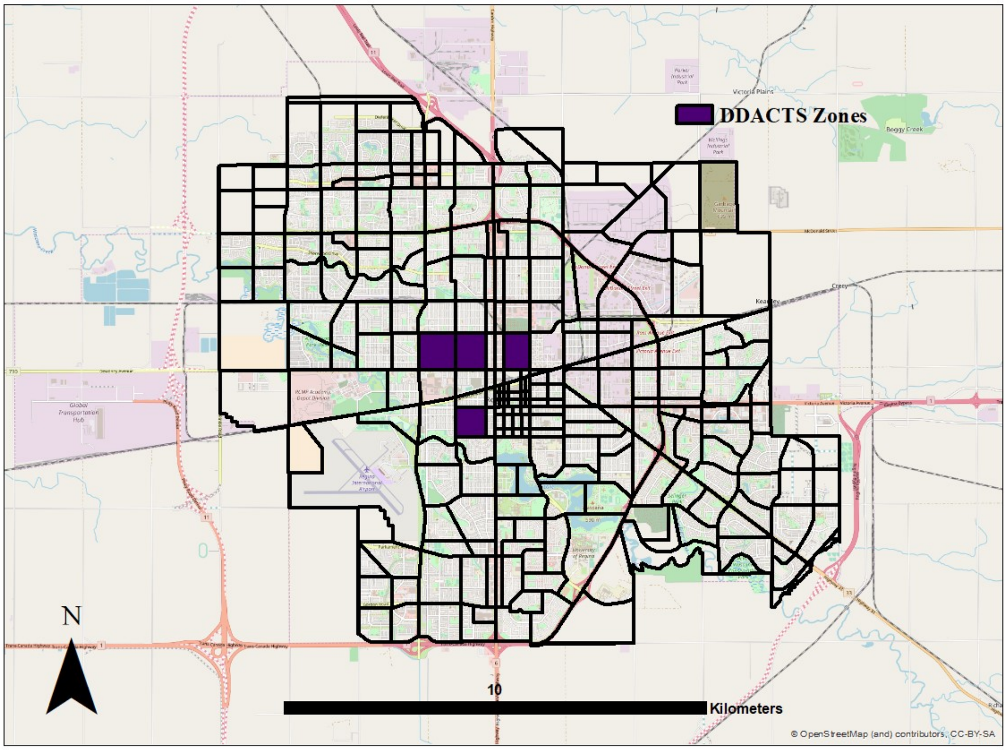


Figure M-8: Non-Violent Crimes and Property Damage Only Collisions

Table M-8: Comparison between the 244 TAZ zones, the Top 10 Hotzones and the 4 DDACTS Zones for Non-Violent Crimes and Property Damage Only Collisions

Categories	Non-Violent Crimes	Property Damage Only Collisions
All Zones numbers of incidences (Crimes and Collisions)	41103 crimes	20883 collisions
Top 10 Zones	8315 crimes	2769 collisions
4 DDACTS hotzones	3885 crimes	1173 collisions
Top 10 hotzones/All 244 Zones	20.23%	13.26%
DDACTS zones/All 244 zones	9.45%	5.62%
Area of all zones	147.93 km^2	147.93 km^2
Area of Top 10 hotzones (km^2)	5.56 km^2	5.92 km^2
Area of 4 DDACTS zones	2.43 km^2	2.43 km^2
Area of Top 10 hotzones/area of all zones	3.76%	4.00%
Area of 4 DDACTS hotzones/area of all zones	1.64%	1.64%