MODELING BICYCLE-VEHICLE CRASH FREQUENCY ON URBAN ROADS

by

Kanya Kamangu Mukoko

A dissertation submitted to the faculty of
The University of North Carolina at Charlotte
in partial fulfillment of the requirements
for the degree of Doctor of Philosophy in
Infrastructure and Environmental Systems

Charlotte

2017

Approved by:

_____
Dr. Srinivas S. Pulugurtha

_____
Dr. Martin R. Kane

_____
Dr. Rajaram Janardhanam

_____
Dr. Wei Fan

_____
Dr. Eric Delmelle

ABSTRACT

KANYA KAMANGU MUKOKO. Modeling bicycle-vehicle crash frequency on urban roads. (Under the direction of DR. SRINIVAS S. PULUGURTHA)

Bicyclists and motorists make mistakes that contribute to traffic crashes involving bicyclists on urban roads. The likelihood of a bicyclist being severely injured or killed daily in traffic crashes is creating fear, anxiety, and becoming a potential danger to the increasing number of Americans using bicycle as a mode of transportation. It is also making bicycling to work or for other purposes less lucrative. Building bicycling friendly and safe environment is, therefore, vital to encourage and have more people use bicycle as a mode of transportation. Therefore, the main goal of this research is to improve safety of bicyclists on urban roads. The main objectives are to understand the role of explanatory variables on risk to bicyclists on urban roads and to develop macroscopic bicycle-vehicle crash frequency models (safety performance functions) for urban roads.

Mecklenburg County in North Carolina was considered as the study area. Reported bicycle-vehicle crash data from 2010 to 2015 along with demographic, land use and network characteristics data was obtained from the local agencies. One-hundred and nineteen locations (intersections) were randomly selected in the study area. These locations were selected such that they are geographically distributed in the study area. Features available in Geographic Information Systems (GIS) software were used to ensure that these locations fall in high, moderate, low and no bicycle-vehicle crash areas. Data within one-mile buffer (vicinity) of 119 randomly selected locations was then captured. These 119 locations accounted for 91.8% of total bicycle-vehicle crashes observed during the study period.

Data for 99 randomly selected locations was used for modeling, while data for the remaining 20 randomly selected locations was used for validating the models. Poisson and Negative Binomial log-link distribution based models were then developed using the modeling dataset. The bicycle-vehicle crash dataset used in this research was observed to be over-dispersed (variance greater than the mean). Therefore, Negative Binomial log-link distribution based models were selected and discussed in this research.

Several demographic, land use and network characteristics were observed to be linearly correlated to bicycle-vehicle crash frequency at a 95% or higher confidence level. Correlations, with p-values = ~0.000, were also observed between demographic, land use and network characteristics (explanatory variables). Six alternate models were developed considering various combinations of explanatory variables, land use and network characteristics, that are not correlated to each other. Two models using all the explanatory variables by ignoring multicollinearity, one each with and without eliminating insignificant explanatory variables, were also developed. The validation dataset was used to compare the estimated bicycle-vehicle crash frequency from each model with the actual bicycle-vehicle crash frequency.

The results obtained from analysis and modeling indicate that bicyclists are at a significantly higher risk of getting involved in a crash while traveling

(1) on segments with no bicycle lane,

(2) on segments with traffic lights,

(3) on segments with 45 mph as speed limit,

(4) in commercial areas,

(5) in areas with research activity and institutions,

(6) in areas with multi-family residential units (densely populated), and,

(7) in heavy industrial areas.

Overall, this dissertation explores interdisciplinary concepts related to transportation engineering, GIS, data analytics and statistical methods to develop and validate models to estimate bicycle-vehicle crash frequency.

ACKNOWLEDGEMENTS

I would like to express my sincere gratitude to my advisor, Professor Srinivas S. Pulugurtha for the continuous support of my Ph.D. study and related research, for his immense encouragement, knowledge, patience, motivation, and wisdom. His guidance helped me with my research and writing of this dissertation. I could not have imagined having a better advisor and mentor for my Ph.D. study.

Besides my advisor, I would like to thank my dissertation committee: Professor Rajaram Janardhanam, Dr. Martin Kane, Dr. Wei fan and Dr. Eric Delmelle, for not only their insightful comments, but also for the hard questions that incented me to widen my research from various perspectives.

Furthermore, I would like to thank Mr. Joe Mangum, AIC., Traffic Safety Coordinator for Charlotte Department of Transportation (CDOT) - Engineering and Operations and Mr. Evan Lowry, GISP., Principal Planner at the Charlotte Mecklenburg Planning department for providing me all necessary data in the requested formats to complete this research.

Last but not least, I would like to thank my family: my wife Esther Mukoko and my daughters Leuticia Mukoko, Christelle Mukoko and Olivia Mukoko and my brother Jean Kanya for supporting me financially and spiritually throughout writing this dissertation and my life in general.

To my daughters: “This is a small work, do more and exceed it”.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

CBD	Central Business District
CDOT	Charlotte Department of Transportation
GIS	Geographic Information Systems
MAD	Mean Absolute Deviation
MAPE	Mean Absolute Percentage Error
MASE	Mean Absolute Scaled Error
MFE	Mean Forecast Error
MGP	Multinomial Generalized Poisson Model
MSE	Mean Square Error
NHTSA	National Highway Traffic Safety Administration
PDO	Property Damage Only
RENB	Random Effects Negative Binomial Model
RMSE	Root Mean Square Error
SMAPE	Symmetric Mean Absolute Percent Error
TRB	Transportation Research Board
Vic Roads	Victoria Roads

CHAPTER 1: INTRODUCTION

A transportation system influences the development of urban areas and serves as an effective way to transport people and goods from one place to another place. It is one of the most basic needs of the human society. Travel and mobility will be very difficult without the provision of an efficient and safe transportation system. While travel is facilitated by the transportation system, mobility is directly influenced by the layout of the transportation network and the level of service it offers to the transportation system users (Beimborn, 1999).

Mobility is a function of urban multilane highways (Wang et al., 2013). On the other hand, accessibility is an important consideration in geometric design and traffic management. An access point, a combination of a median opening and its served minor roads within an urban multilane road section, is bounded by two consecutive signalized intersections (critical spots that influence the safety and mobility of urban multilane highways) with turning movements (left-turn, through and right-turn) and/or U-turn movements. As population and traffic volume increase on urban multilane roads, interactions between vehicles, pedestrians and bicyclists on roads become more frequent and serious. Due to high vehicle speed, high traffic volume, and high access movements, pedestrians and bicyclists are the most vulnerable users of urban multilane highways (Wang et al., 2013).

Urban settings world-wide are ideal for bicycling to become a significant mode of transportation, given the greater compactness of destinations (Delmelle and Thill, 2008;

Wei and Lovegrove, 2012). Chaurand and Delhomme (2013) acknowledged that today's increase in the number of bicyclists has triggered a change in the interactions to be handled by transportation system users. Undoubtedly, increasing the number of bicyclists on urban streets is a sustainable solution to congestion and air quality problems encountered in most cities. Indeed, ecological issues concerning the environmental consequences of the use of motorized transportation, concerns about the impact of car use on health, or problems in terms of cost and time loss due to traffic congestion are leading people to change their transportation mode. The above dynamic is encouraged by public policies, through campaigns in favor of sustainable transportation (Chaurand and Delhomme, 2013). However, in the United States, bicycling is both scarcely used and very dangerous, as bicyclists are 12 times more likely to be killed in a road crash than motorists (Delmelle and Thill, 2008). They have a higher risk of being injured in a road crash compared to other road users (Martinez-Ruiz et al., 2013).

Bicyclist's safety is a major concern to urban transportation planners, engineers and system managers (LaMondia and Duthie, 2012; Wei and Lovegrove, 2012; Martinez-Ruiz et al., 2013). In a survey of bicyclists in Texas, 69% of the respondents stated that they feel bicycling is "somewhat dangerous" or "very dangerous" from the standpoint of traffic crashes (LaMondia and Duthie, 2012). However, relatively fewer number of studies were conducted on crash risk perceived by bicyclists interacting with other transportation system users, bicyclists' and motorists' perceptions of crash risk in bicycle–vehicle interactions, or modeling to estimate bicycle-vehicle crash frequency on urban roads.

1.1 Problem Statement

Pedalcyclists (example, Figure 1) are bicyclists and other cyclists including riders of two-wheel, non-motorized vehicles, tricycles, and unicycles powered solely by pedals (NHTSA, 2015). A traffic crash (example, Figure 2 and Figure 3) is defined as an incident that involved one or more vehicles where at least one vehicle was in transport and the crash originated on a public traffic way, such as a road or highway (NHTSA, 2015). Crashes that occurred on private property, including parking lots and driveways, are typically excluded. Further, bicycle-vehicle crashes, as per the National Highway Traffic Safety Administration (NHTSA)'s fact sheet, do not include bicycle wrecks that do not involve vehicles.

Figure 1: Bicyclist

Figure 2: Bicyclist crash

Figure 3: Traffic crash

The NHTSA's ("Safety in Numbers Newsletter" titled "Bicycles") acknowledged that from 2000 to 2012, the number of Americans traveling to work by bicycle increased from ~488,000 to ~786,000 (U.S. Census Bureau, May 2014).

While a bicycle can offer many health, financial, and environmental benefits, it can also bring the dangers associated with any vehicle. Bicyclists and motorists make mistakes

that contribute to crashes. When a crash happens involving a bicyclist and a car, SUV, pickup truck, or bus, it is the bicyclist who is likely to be injured or killed. Of those injuries, a significant number are incapacitating, meaning the bicyclist could not leave the crash scene without assistance (skull, chest, or abdominal injuries, broken limbs, severe lacerations, or unconsciousness) (NHTSA, 2014).

In the United States, 743 bicyclists were killed and an estimated 48,000 injured in traffic crashes in 2013 (NHTSA, 2015). Bicyclist deaths accounted for 2% of all traffic fatalities during the year. The number of bicyclists killed in 2013 is 1% higher than the 734 bicyclists killed in 2012. The increase in 2013 is the third straight increase in bicyclist fatalities, a 19% increase since 2010 (NHTSA, 2015). While 68% of bicyclists were killed in urban traffic crashes in 2013, 57% of bicycle-vehicle crashes were recorded at non-intersections.

The likelihood of a bicyclist being severely injured or killed daily in traffic crashes is creating fear, anxiety, and becoming a potential danger to the increasing number of Americans considering bicycle as a mode of transportation. The number of incapacitating injuries to bicyclists is dramatic and terrifying for Americans in need for healthy, financial, and environmental friendly alternate modes of transportation (NHTSA, 2014). There is a need to better understand the role associated factors and proactively plan to reduce bicycle-vehicle crash frequency on urban roads.

The lack of physical exercise has led to higher rates of obesity, hypertension and diabetes among Americans compared to most Europeans. Europeans have longer healthy life expectancies, although they spend less than half as much as Americans on healthcare because of their overall physical exercise levels, primarily attributed to much higher rates

of walking and bicycling (Pucher and Dijkstra, 2003). Pucher and Dijkstra (2003) encouraged promoting safe and convenient walking and bicycling for daily urban travel to improve public health.

For a nation to remain strong and find its rightful place in the community of nations, significant investment in bicycling infrastructure to improve bicycling conditions is required through the three "pillars" (social, economic, and environmental) of sustainable transportation systems and modes (Figliozzi et al., 2013). The sustainability of the transportation modes is very important as it accounts for the social (improved health due to physical activity and other related reasons), the economic (cost efficiency), and the environmental (less congestion, no fossil fuel consumption, no air or noise pollution) spheres (Farley and Smith, 2014).

Sustainable transportation systems, including walking, bicycling, public transit, green vehicles, and car sharing, make more positive contributions to the society, the economy and the environment than automobile dominated transportation systems (Wei and Lovegrove, 2012). The benefits of bicycling generally include relatively low costs, emissions, and energy use, together with improved health, and convenient parking. Bicycling remains one of the most effective modes for short trips with distances less than 3.1 miles. In fact, bicycling is typically the fastest mode for trips less than 3.1 miles. This transportation mode is being encouraged for use widely. However, as vulnerable road users, bicyclists are more likely to be injured when involved in crashes (Wei and Lovegrove, 2012).

LaMondia and Duthie (2012) indicated that the transportation community has long been divided over the most appropriate and safest way to accommodate bicyclists and

motorists together on multilane roadways. The alternatives range from separated pathways to designated bicycle lanes to requiring bicyclists to share the road. Regardless of which alternative is correct, or if a correct position exists, facilities for bicyclists are not abundant and, thus, space is shared for at least some segments of most trips (LaMondia and Duthie, 2012).

Traffic engineers, professionals, practitioners, policy makers and authorities are, therefore, obligated to bring improvements and assure that bicyclists' safety is provided and maintained on urban roads. Further, NHTSA recommends effective actions that communities can take to improve bicycle safety for adults and children (NHTSA, 2014).

1.2 Research Goal and Objectives

Common modes of transportation include bicycles, cars, buses, other vehicles, trains, boats and planes. Transportation projects are planned to decrease travel time and improve safety irrespective of the mode of transportation. Despite benefits provided by the transportation system, negative developments arising from it cannot be ignored. Fatal, injury and property damage only (PDO) crashes are typical spin-offs from the transportation system.

Roess et al. (2004) state that traffic engineering is a phase of transportation engineering which deals with the planning, geometric design and traffic operations of roads, streets, highways, their networks, terminals, and abutting lands. In traffic operations, safety is one of the primary objective. Therefore, the provision of a safe transportation system is an important responsibility of a traffic engineer.

From NHTSA's 2012 national representative telephone survey, a typical day average duration of a bicycle ride is about 45 minutes. The most common ride length is

30 minutes or less (42%). The most commonly cited purpose of bicycle trips is recreation (33%) and exercise (28%), followed by personal errands (17%), visiting a friend (8%), commuting to work (7%) or going to school (4%). More fatalities occurred during the summer months of July through September (NHTSA, 2014).

The pedestrians and bicyclists traveling to school have the highest rate of injury and fatality on a per-mile basis (McMillan, 2007). Engineering, enforcement, and education, the "3 E's", are critical and needed to reduce crashes and save bicyclist lives. In summary, bicyclists are one of the most vulnerable class of transportation system users. Therefore, the main goal of this research is to contribute and improve the safety of bicyclists on urban roads.

Despite numerous efforts on the safety of bicyclists, there are still some unanswered research questions. They are:

(1) What is the relationship between various demographic (population, household units, and household mean-income), land use (business, business park, business district, mixed use, mixed use residential, light industrial, heavy industrial, manufactured home, single-family, multi-family, institutional, research, commercial, office, transit oriented development, uptown mixed use, and urban residential) and network (number intersections by control type, number of bus-stops, number of elementary schools, number of middle schools, number of high schools, number of private schools, number of colleges and universities, center-line miles with bicycle / no bicycle lane, center-line miles with sidewalk / no sidewalk, center-line miles of divided / undivided roads, center-line miles by number of lanes, and center-line

miles by speed limit) characteristics and bicycle-vehicle crashes in the vicinity of a location (intersection)?

(2) Do they have a positive or negative influence on bicycle-vehicle crash frequency?

(3) How can one model and estimate bicycle-vehicle crash frequency within the vicinity of a location?

(4) Would eliminating correlated and statistically insignificant explanatory variables enhance model's predictability?

(5) Are such models valid to estimate bicycle-vehicle crash frequency and proactively identify and implement potential countermeasures.

Thus, the following research objectives were identified and selected to accomplish the goal.

1. Apply Geographic Information Systems (GIS) tools to identify a geographically distributed, unbiased sample of locations in high, medium, low and no risk bicycle-vehicle crash locations.

2. Capture, analyze and understand the role of explanatory variables on bicycle-vehicle crashes on urban roads.

3. Examine correlations and select explanatory variables to develop bicycle-vehicle crash frequency models.

4. Develop and validate macroscopic bicycle-vehicle crash frequency models for urban roads.

It is hypothesized that surrogate data such as demographic, land use and network characteristics within the vicinity of a location can be used to model and estimate bicycle-

vehicle crashes. While characteristics or explanatory variables in each category (demographic, land use and network) may influence bicycle-vehicle crashes, network characteristics may be better predictors of bicycle-vehicle crash frequency. Further, the use of selected, significant but not correlated, explanatory variables may yield similar or better statistically meaningful outcomes.

It is envisioned that the models and valid estimates can be proactively used to develop comprehensive transportation plans, metropolitan transportation plans, and transportation improvement programs as well as assist with land use decisions. The findings from this research are targeted towards planners, professionals, practitioners and policy-makers who might be able to correct the hazards and improve safety of bicyclists on urban roads. In conjunction with policy-makers, the "3 E's" could be used to further minimize crashes involving bicyclist's in the United States cities as well as around the world.

1.3 Dissertation Outline

This dissertation is organized as follows. Chapter 2 presents a review of past studies and researches on identification of crash locations and data extraction, bicycle-vehicle crash risk factors, overview of crash injury severity and crash frequency modeling, and limitations of past research. Chapter 3 discusses the data required and outlines the methodology adopted in this research. Chapter 4 presents the selection of locations for data extraction, correlation between explanatory variables and the selection of explanatory variables to develop bicycle-vehicle crash frequency models. Chapter 5 describes the

bicycle-vehicle crash estimation models and results from model validation. Chapter 6 presents conclusions from this research and scope for future work.

CHAPTER 2: LITERATURE REVIEW

The discussion of past literature is presented in this chapter. The chapter is divided into four sections. Identification of crash locations by risk level and data extraction are discussed in Section 2.1. Factors contributing to bicycle-vehicle crashes are described in Section 2.2. A review of various methods used in modeling crash injury severity and traffic crash frequency is presented in Section 2.3. Limitations of past research are discussed in Section 2.3.

2.1 Identification of Crash Locations and Geospatial Data Extraction

Several researchers have studied, developed methods and extracted geospatial data to analyze transportation problems. The section is divided into two subsections. Literature pertaining to identification of crash locations is discussed in the first subsection, while literature pertaining to geospatial data extraction to analyze transportation problems is discussed in the second subsection.

2.1.1 Identification of Crash Locations

Identifying traffic crash locations is very important prior to allocation of resources and determining effective strategies for the reduction of crashes. However, literature provides no universally accepted definition of a traffic crash "hotspot" (Anderson, 2009). Dealing with the presence of several crashes at one point location, Pulugurtha et al. (2005) warned that in a spatial distribution of pedestrian crashes, the presence of a dot does not necessarily equal one crash. Several crashes may have occurred at the point. Therefore, it is difficult to identify locations that have multiple crashes using a spatial distribution map.

The suggested solution for this problem is crash density or concentration maps (Pulugurtha et al., 2005). The development of such maps works by grouping point features within a certain distance of one another into one symbol, while reducing the complexity within the dataset (Delmelle, 2016).

The density analysis takes geocoded crash data and spreads them across the study area based on the quantity that is measured at each location (say, risk or number of crashes per unit area) and the spatial relationship of the locations of the measured quantities. The resulting surfaces surrounding each point in Kernel density are based on a quadratic formula with the highest value at the location of a crash. The results obtained from identification and ranking of crash locations are sensitive to buffer radius, cell size and the ranking methods (Pulugurtha and Vanapalli, 2008).

Several researchers explored the concept to accomplish the task of identifying high crash locations. Pulugurtha and Nambisan (2003), Pulugurtha et al. (2005) and Pulugurtha et al. (2007) identified high pedestrian crash locations and defined criteria to rank (prioritize) them by examining spatial clustering and dispersion of pedestrian crashes. Delmelle and Thill (2008) and Delmelle et al. (2008) used geospatial methods to determine the geographic distribution of crashes and crash hazard intensity factors for both youth and adult urban bicyclists in Buffalo, New York. Pulugurtha and Vanapalli (2008) and Pulugurtha and Penkey (2010) identified pedestrian crashes on segments with transit service, pedestrian crashes on segments without transit service, unsafe segments for transit system users who walk to access the system, and hazardous bus-stops using pedestrian crash data.

Other related geospatial methods adopted in the past include (1) the use of Anselin Moran's Local Spatial Autocorrelation tool to detect high crash clusters and identify factors that influence the concentration of pedestrian crashes (Flahaut et al., 2003; Emaasit et al., 2013); (2) the use Kernel Density Estimation (KDE) to study the spatial patterns of injury related traffic crashes and to create a classification of traffic crash locations (Anderson, 2009); (3) the use of network K-functions in traffic accident analysis when compared to planar K-functions (Yamada and Thill, 2004); and, (4) the use spatial KDE to examine the clustering patterns of pedestrian crashes and identify locations where clustering is more pronounced (Jang et al., 2013).

The criteria to define high crash locations is also an important factor for allocation of resources for safety improvements. Pulugurtha and Nambisan (2003) considered high risk locations (zones) as those with a target rate of 10 pedestrian crashes per zone during the study period, compared to a total annual size of 200 pedestrian crashes in the study area. Armstrong and Petch (2013) classified any location with two or more pedestrian crashes in the previous three years to be a high pedestrian risk location. Pulugurtha and Imran (2013) categorized signalized intersections based on risk to pedestrians and bicyclists using Jenks natural breaks inherent in the data. However, categorizing locations based on bicycle-vehicle crash data is difficult due to relatively fewer number of bicycle-vehicle crashes observed in urban areas annually.

2.1.2 Geospatial Data Extraction and Transportation Problems

Literature documents several efforts to extract geospatial data and analyze transportation problems. As an example, features available in GIS software were used (1) to extract data and identify critical factors for modeling pedestrian activity at signalized

intersections (Pulugurtha and Repaka, 2008); (2) to extract data and develop models to estimate pedestrian demand by the level of pedestrian activity at signalized intersections (Pulugurtha and Repaka, 2011); (3) to extract data and examine the role of the number of bus-stops and factors such as demographic, socio-economic, land use, and network characteristics on transit ridership (Pulugurtha and Agurla, 2012); (4) to extract census and land use data for developing models to estimate crashes at intersections- with and without using traffic volumes (Pulugurtha and Nujjetty, 2012); and, (5) to extract geospatial data pertaining to low-income communities and examine their influence on the number of crashes compared to other areas (Kravetz and Noland, 2012).

The locations or segments for data extraction, analysis and modeling should be randomly selected to avoid any bias in assessing the relationship, ranking, and allocation of resources for transportation improvements (Pulugurtha and Penkey, 2010). The width of the buffer also plays a vital role in extracting geospatial data for analysis and modeling. Bolstad (2012) defined buffer as a region that is less than or equal to a specified distance from one or more features.

Hess et al. (1999) studied the relationship between site design and pedestrian travel in mixed-use and medium density environment. Geospatial data was captured within a 0.5-mile pedestrian travel catchment area. Their study defined commercial-center size using the number of businesses and types of retail facilities provided within the 0.5-mile pedestrian catchment area (Hess et al., 1999).

The pedestrians at risk of getting involved in a crash was estimated by generating both Euclidean and network buffers (circular and linear zones) of width equal to accessible walking distance (say, 0.5-mile) around each location, and estimating the population

residing in the vicinity of each location (Pulugurtha and Nambisan, 2003). Falb et al. (2007) used GIS to estimate the percentage of potential walkers (school-age children) living within 1-mile from public schools in Georgia. The 1-mile study area was called as "pedestrian catchment area" (Falb et al., 2007).

McDonald (2008) indicated that, in general, two factors have an impact on active transportation: (1) individual/household factors (age, gender, race, household income and vehicle availability); and (2) neighborhood factors (population density and neighborhood disadvantage). Their study revealed that living within a 0.5-mile of school greatly increased the likelihood of walking or bicycling to school across all groups. Also, rates of active transportation varied significantly by racial/ethnic and income groups (McDonald, 2008).

Pulugurtha and Repaka (2008, 2011) used different buffer widths (proximal area) to extract network characteristics and off-network (demographic and land use) characteristics to estimate pedestrian counts. Likewise, Zahabi et al. (2011) generated buffers of five different sizes (0.0310-, 0.0621-, 0.0932-, 0.1242-, and 0.2485-mile) to capture the urban form and environment variables that have an effect on the severity of the crash instead of using a single buffer. Each crash record was georeferenced to its exact coordinates and the built environment measures were computed to capture the context of the area where it occurred (Zahabi et al., 2011).

Pulugurtha and Agurla (2012a,b) used different buffers (0.25-, 0.5-, 0.75 and 1-mile) to evaluate the best proximity distance that has a strong influence on pedestrian and transit activity at a bus-stop. Network characteristics from aerial photographs and field visits were then added to the databases. Demographic, socio-economic and land use

characteristics around each bus-stop were overlaid on generated buffers to extract data and develop models to estimate pedestrian activity (counts) and transit ridership at bus-stops. Distance decay effect was also adopted to integrate data from the different buffers and estimate pedestrian activity and transit ridership at bus-stops. However, their research did not show that using integrated data from different buffers would yield significantly better results than from individual buffers.

Emaasit et al. (2013) overlaid generated clusters on selected socio-economic and population data to examine their association with high crash clusters. The minimum distance to ensure every crash incident has at least one neighbor was determined to be 3.4 miles (Emaasit et al., 2013).

One-mile buffer around a location might be the best proximal distance to extract geospatial data and model bicycle-vehicle crashes for two reasons: (1) distance (decision to bicycle depends on how long is the travel distance); and (2) travel time (takes about 5 to 6 minutes for a bicyclist at an average speed to cover 1-mile distance (Wei and Lovegrove, 2012). Further, the average distance for bicyclists aged 15 years and under is ~0.58 miles, while the average distance for those bicyclists older than 15 years is ~1.15 miles. Younger bicyclists are, however, unlikely to travel far from their own neighborhoods (Delmelle et al., 2008).

2.2 Bicycle-Vehicle Crash Risk Factors and Modeling

Literature documents research on factors contributing to bicycle-vehicle crashes, risk to bicyclists and analyzing bicycle-vehicle crashes at traffic analysis zone (TAZ) level.

Wachtel and Lewiston (1994) compared age, gender, direction of travel (with or against traffic flow), and position on the road (in roadway, bicycle lanes, private driveways,

sidewalk, paths or crosswalks) of bicyclists involved in crashes with similar data for the general population of bicyclists observed along the same streets. The comparison enabled to identify factors that are correlated with increased risk of bicycle-vehicle crashes. Other variables that increase the risk of bicycle-vehicle crashes are urban areas, intersections, land use (residential, business district, etc.), road cross-section (two-, four- or six-lane), annual average daily traffic (AADT) and posted speed limit. They warned that intersections (interpreted broadly) are the major point of conflict between bicycles and vehicles. Sidewalk for bicycling, adjacent to busy streets with many intersections, presents special dangers and should not be encouraged through the construction or designation of bicycle paths parallel to the street (Wachtel and Lewiston, 1994).

Klop and Khattak (1999) examined the effect of physical and environmental factors on injury severity in bicycle-vehicle crashes. Over the four years (1990-1993) examined in bicycle-vehicle crashes on state-controlled two-lane undivided roads in North Carolina, 60 bicyclists were killed and 947 bicyclists were injured in police reported bicycle-vehicle crashes. Their study indicated that road cross-section elements (two-lane) and environmental factors as well as individual, vehicle, and bicycle factors drive the crash process. Reaction times of both the motorist and the bicyclist, perceptual and judgement errors, and attention also affect bicycle-vehicle crash frequency. In addition, motorist and bicyclist information processing, motorist and bicyclist behaviors, the care that some motorists use when near bicyclists, and right-turn-on-red situations may also increase or decrease injury severity. Variables expected to significantly influence injury severity among bicyclists are curves, upgrades, downgrades, intersections, driveways, alleys,

parking lots, narrower lanes (width), shoulder width, increasing AADT, speed limit, street lighting, age-group, rain and fog (Klop and Khattak, 1999).

Kim et al. (2006) explored the factors contributing to the injury severity of bicyclists in bicycle-vehicle crashes using a multinomial logit model. The analysis is based on police reported crash data from 1997 to 2002 for North Carolina. They predicted the probability of four injury severity outcomes: fatal, incapacitating, non-incapacitating, and possible or no injury. Their study included demographic and economic characteristics (age, gender, ethnicity, and income), road characteristics (speed limit, intersections, type of roads, and pavement), road geometry (curved, straight, and grade), environmental factors (month, day, time, weather, and road surface), locations or land characteristics (urban or rural areas, driveway, shoulder, bicycle lane, and trail), land use characteristics (residential, institutional, industrial, and commercial), direction of travel, the influence of alcohol, head injuries, bicycle helmet usage, crash types, and party at fault. Socio-economic factors, particularly the percentage of poor households within a neighborhood, played an important role in the prediction of bicycle crash rates. Family and neighborhood characteristics were stronger risk factors for bicycle injuries than children's personality and behavior. Higher risk of injury in children was related to fewer years of parent education, a history of crashes in the family, an environment judged as unsafe, and poor parental supervision (Kim et al., 2006).

Density of development, physical road characteristics (roadway and intersection), socio-economic and demographic variables, and potential trip attractors were examined using Buffalo, New York bicycle crash data (Delmelle and Thill, 2008). In another study, bicycle crashes were analyzed to determine and compare risk factors of both child and adult

bicyclists (Delmelle et al., 2008). In a recent study, it was found that child bicyclists (<10 years) are more likely to be involved in non-intersection bicycle-vehicle crashes (Hamann et al., 2015).

Reynolds et al. (2009) reviewed studies of the impact of transportation infrastructure on bicyclist safety. The results were tabulated within two categories of infrastructure at intersections (e.g., roundabouts, traffic lights, bicycle crossings or intersection design) and between intersections on "straightaways" (e.g., bicycle lanes or paths, road design characteristics, road surface, sidewalk and street lighting). Their study found that multilane roundabouts can significantly increase risk to bicyclists, unless a separated bicycle track is included in the design. Sidewalks and multi-use trails pose the highest risk to bicyclists. The major roads are more hazardous than minor roads for bicyclists. The presence of bicycle facilities (e.g., on-road bicycle routes, on-road marked bicycle lanes, and off-road bicycle paths) was associated with the lowest risk to bicyclists (Reynolds et al., 2009).

Zahabi et al. (2011) investigated the link between built environment characteristics and pedestrian-vehicle and bicyclist-vehicle crash severity. They believed that road facilities in urban areas are a major source of injury for non-motorized road users despite the benefits of non-motorized transportation. The location, road design, and urban form characteristics influence pedestrian-vehicle and bicycle-vehicle crashes and the severity of the injury sustained in the crashes (Zahabi et al., 2011). Kravetz and Noland (2012) used spatial autocorrelation and Negative Binomial distribution to analyze to what extent crash disparities occur due to inequitable road infrastructure, and whether this disparity can be linked to the socio-political disparities in the region (Kravetz and Noland, 2012).

Moore et al. (2011) developed multinomial logit and mixed logit models to estimate the degree of influence that bicyclist, driver, motor vehicle, geometric, environmental, and crash type characteristics have on bicyclist injury severity. They observed that factors affecting bicyclist injury severity at intersection and non-intersection locations are substantively different. Stipancic et al. (2015) developed a segmented ordered logit model for bicycle-vehicle conflict occurrence to evaluate the impact of gender on bicyclist risk at urban intersections with bicycle lanes. They found that male bicyclists, with all else being equal, are less likely to be involved in conflicts than female bicyclists.

Delmelle et al. (2012) studied the relative risk factors of bicycle and pedestrian crashes at the neighborhood level using data for the city of Buffalo, NY. Their analysis underscored significant differences tied to neighborhood ethnicity, educational attainment and land use, while physical characteristics of the road infrastructure were registered as marginally discriminating factors. Income related socio-economic status was not found to play a prominent role in bicycle-vehicle crashes (Delmelle et al., 2012). Contrarily, Hamann et al. (2015) observed that bicycle-vehicle crashes occur more frequently in low income and education areas.

LaMondia and Duthie (2012) studied the impact of roadway environment, motorist behavior, and bicyclist behavior on bicyclist-motorist interactions. Their study considered three distinct components: bicyclist lateral location, bicyclist-motorist interaction movement, and bicyclist-motorist lateral interaction distance. Each of these components provide insight into a specific, but related aspect of how bicyclists and motorists relate to each other on a road. Three unique ordered probit regression models that describe bicyclist

lateral location, bicyclist-motorist interaction movement, and bicyclist-motorist lateral interaction distance were developed (LaMondia and Duthie, 2012).

Wei and Lovegrove (2012) revealed that an increase in bicycle–vehicle crashes is associated with an increase in total lane miles, bicycle lane miles, bus-stops, traffic signals, intersection density, and arterial–local intersection percentage. Models were categorized in four groups: (1) urban exposure (total lane miles, total bicycle lane miles, and zonal area); (2) urban socio-economic and demographics (population density, home density, employed density, and average income); (3) urban transportation demand management or network (commuter density, core area, transit commuter, bicycling commuter, pedestrian, and bus-stop density); and (4) urban road network (signal density, intersection density, and number of arterial-local intersections) (Wei and Lovegrove, 2012).

Martinez-Ruiz et al. (2013) used the choice of exposure metrics approach to identify motorist-related and vehicle-related factors associated with the risk of causing a road crash involving a bicyclist in Spain. The method constitutes a potentially useful tool that compares the characteristics of responsible and non-responsible motorists involved in road crashes. They retained the following risk factors for causing road crashes: age group, gender, psychophysical circumstances (DWI), helmet use, hours driving without a rest, type of bicyclist (professional or not), bicyclist maneuver before crash (passing, turning, crossing intersection, etc.), number of occupants (one or more than one), and bicycle defects (lights, brakes) (Martinez-Ruiz et al., 2013).

Chaurand and Delhomme (2013) studied bicyclists' and motorists' perceived risk in bicycle-vehicle interactions. Bicyclists' presence on the road is considered annoying by motorists and even regarded as a source of danger. As such, majority of motorists refuse

to use a bicycle as their transportation mode because of the risk they feel they would run when riding among cars and other motorized vehicles, and also because they perceive cars as a protective, safe “cocoon” (Chaurand and Delhomme, 2013).

Zhang et al. (2013) analyzed the associations between road network structure and pedestrian-bicyclist crashes and identified relationships between dependent and explanatory variables across locations. The dependent variable is the average number of crashes involving pedestrian and bicyclist per year. The explanatory variables were classified into five categories: structural measures, land use, travel behavior, transportation facilities, and demographic features (Zhang et al., 2013).

Wang et al. (2013) categorized access designs (variables) into six types based on the design of median openings and the number of legs. Access Type I (three-leg access point with closed median opening), Access Type II (three-leg access point with directional median opening), Access Type III (three-leg access point with full median opening), Access Type IV (four-leg access point with closed median opening), Access Type V (four-leg access point with directional median opening), and Access Type VI (four-leg access point with closed median opening) have an effect on the occurrence of pedestrian-bicycle crashes (Wang et al., 2013).

While bicyclist numbers continue to rise and the benefits continue to be enjoyed, bicycling in urban environments still comes with serious safety concerns, in particular, at intersections (Strauss et al., 2013; Figliozzi et al., 2013). They observed that majority of bicyclist injuries occur at intersections. Bicycling activity through intersections was found to increase as employment, number of metro stations, land use mix, area of commercial land use type, length of bicycle facilities, and the presence of schools within 0.031–0.497

mile of the intersection increase. Intersections with three approaches are expected to have fewer bicyclists than those with four approaches. The expected injury frequency and injury rates were estimated for each intersection and used to rank corridors (Strauss et al., 2013).

Pulugurtha and Imran (2013) overlaid pedestrian-vehicle and bicycle-vehicle crash data, separately, on Kernel density maps to examine the spatial relation between pedestrian and bicycle level of service (LOS) and crash locations. They examined the effect of the distance from downtown/uptown on pedestrian and bicycle LOS (Pulugurtha and Imran, 2013). Chimba et al. (2014) identified patterns of pedestrian-vehicle and bicycle-vehicle high crash locations and flagged combination of demographic, socioeconomic and geometry variables that are good indicators of areas likely to experience pedestrian-vehicle and bicycle-vehicle crashes.

Nordback et al. (2014) developed safety performance functions for bicyclists to serve as a basis for future investigations and for prioritizing intersections to better allocate scarce funds for bicycle safety improvements. They found that intersections with higher bicyclist counts and higher motorist counts have higher bicycle-vehicle collisions.

Pulugurtha and Thakur (2015) evaluated the effectiveness of on-street bicycle lane (in reducing crashes involving bicyclists) and assessed the role of network characteristics (number of lanes, speed limit, etc.) on risk to bicyclists on urban roads. The results obtained from their analysis indicated that bicyclists are three to four times at higher risk (based on traffic conditions) on segments without on-street bicycle lane than when compared to segments with on-street bicycle lane. Bicyclists are also more susceptible to crashes on high speed / high traffic volume multilane roads (Pulugurtha and Thakur, 2015).

Kaplan and Giocomo Prato (2015) researched to unravel land use and network factors contributing to the probability of being involved in a crash, conditional on the crash occurrence, experiencing a severe injury outcome. Wang et al. (2015) investigated factors correlated with the severity of injuries sustained by bicyclists in bicycle-vehicle crashes at unsignalized intersections to develop site-specific countermeasures and interventions to improve bicyclist safety. They found that stop-controlled intersections, one-lane approaches, helmet usage, and lower speed limits were associated with decreased injury severity, while uncontrolled intersections, inadequate lighting condition, and wet road surfaces increased injury severity.

Amoh-Gyimah et al. (2016) researched on various factors that influence the occurrence of pedestrian and bicycle crashes at the planning level. They found that vehicle miles traveled, percent of old population, percentage of households without vehicles have a significant and positive correlation with the number of pedestrian and bicycle crashes.

A few researchers focused on analysis and modeling of bicycle-vehicle crashes at TAZ level. Wei and Lovegrove (2012) extracted demographic and network data at TAZ level to evaluate the safety of bicyclists. Nashad et al. (2016) conducted statewide TAZ level analysis to allow planners identify high-risk zones for pedestrians and bicyclists, for screening and subsequent treatment identification. The role of factors such as accessibility measures, exposure measures, demographic characteristics and network characteristics on bicycle-vehicle crashes at TAZ level was also researched (Yasmin and Eluru, 2016).

2.3 Crash Injury Severity and Crash Frequency Modeling

This section is divided into two subsections. The first subsection focuses on discrete choice modeling, while the second subsection focuses on count regression modeling.

2.3.1 Discrete Choice Modeling

Discrete choice modeling is a statistical procedure to model choices made by people among a finite set of alternatives (Fils, 2012). The procedure has been used to examine the choice of which car to buy, where to go to college, which mode of transportation to take to work, and the number of vehicles a household chooses to own. While regression analysis examines "how much", discrete choice analysis examines "which". It is important that the potential outcomes must be discrete i.e., if a response variable *Y* is binary, it can have only two possible outcomes, 1 or 0 (Fils, 2012). The binary and multinomial models are the most common discrete choice models. The only distinction between them is that a binary model considers two discrete outcomes, while a multinomial model considers three or more discrete outcomes. These alternative models are widely used and applied in many transportation data analysis as alternatives to linear regression modeling (Fils, 2012).

Several researchers used discrete choice modeling to assess crash injury risk in the past. O'Donnell and Connor (1996) compared ordered logit and assessed the probabilities of four levels of injury severity as a function of motorists' attributes. Abdel-aty et al. (1998) used multinomial logit model to examine relationships between motorist age and crash characteristics. Zahabi et al. (2011) developed injury severity models to investigate the effect of road design, built environment, speed limit, and other factors (e.g., vehicle characteristics and movement type) on injury severity levels of pedestrians and bicyclists

involved in crashes with vehicles. An ordered logit model was used to estimate the effects of each of the variables on the severity of the injury sustained in a crash; because the injury severity is ordinal in nature (Zahabi et al., 2011).

LaMondia and Duthie (2012) used a unique methodology for studying bicycle–vehicle interactions through the use of three distinct, unique ordered probit regression models that describe the three interaction components. These models predict interaction choices (e.g., where within the road a bicyclist will choose to travel), as well as identify those factors that influence these choices. This structure of discrete choice model was used to assess bicyclists' and motorists' mutual acceptance and comfort level sharing a road (LaMondia and Duthie, 2012).

Chiou and Fu (2013) developed a multinomial generalized Poisson (MGP) model to simultaneously model crash frequency (count data) and severity (ratio data). MGP model is an extension of the multinomial-Poisson regression model and assumes that crashes can be classified into a finite number of clusters according to severity levels. Also, the frequency of each severity level follows a conditional multinomial distribution.

Bin Islam and Hernandez (2013) developed a random parameter logit model to predict the likelihood of five standard injury severity (KABCO) scales commonly used in the Crash Records Information System (where, K = fatal, A = incapacitating injury, B = non-incapacitating injury, C = possible injury, and O = PDO). Contributing factors considered include motorist demographic characteristics, traffic flow, road geometric features, land use characteristics, time characteristics, weather, and lighting conditions.

Fan et al. (2015a, 2015b) developed a multinomial logit model using SAS PROC LOGISTICS procedure. The three pedestrian crash severity levels (fatality, injury and no

injury) were considered as dependent variables. Pedestrian characteristics, environmental factors, type of development area, highway-rail crossing characteristics, highway traffic characteristics and train speed were considered as the explanatory variables when predicting crash severity levels.

2.3.2 Count Regression Modeling

A common mistake is to model count data as continuous data by applying standard least squares regression, which is not correct because regression models yield predicted values that are non-integers. Also, regression can predict values that are negative, which together with non-integers are inconsistent with count data (Anastasopoulos et al., 2008). These limitations make standard regression analysis inappropriate for modeling count data without modifying dependent variables (Washington et al., 2003; Fils, 2012).

Wang et al. (2015) stated that count models are commonly applied for road segments, intersections and TAZs to identify factors related to the occurrence of crashes. The most frequently applied distributions to develop count models include Negative Binomial and Poisson models, zero-inflated Poisson and zero-inflated Negative Binomial distribution based models, and random parameter Negative Binomial distribution based models.

Considering the discrete, sporadic, and random characteristics of crash data, the Poisson distribution based models appear to be suitable and have been used by many researchers (Ma et al., 2015). As an example, Ivan et al. (2000) developed Poisson regression models to estimate single and multi-vehicle crash rates as a function of traffic density, land use, ambient light conditions and time of day.

The limitation of the Poisson model is that the mean must be equal to the variance. In fact, many researchers found that the variance is much greater than the mean, which indicates that crash data may be over-dispersed (Miaou, 1994; Shankar et al., 1995; Vogt and Bared, 1998). To overcome the problem of over-dispersion, researchers have applied the Negative Binomial distribution based model instead of the Poisson distribution based model (Miaou, 1994; Shankar et al., 1995; Poch and Mannering, 1996; Abdel-Aty and Radwan, 2000).

A few example studies related to Negative Binomial distribution based crash count models are outlined next. Poch and Mannering (1996) developed a Negative Binomial distribution based crash frequency model for intersection approaches. Pulugurtha and Nujjetty (2012) developed Negative Binomial distribution based count models (to account for observed over-dispersion) to estimate the number of crashes at intersections for two different scenarios. While models were developed considering all variables (including traffic volume) that are not correlated to each other as explanatory variables in the first scenario, models were developed considering all variables (excluding traffic volume) that are not correlated to each other as explanatory variables in the second scenario. The numbers of crashes at each intersection was used as a dependent variable. Demographic, socio-economic, and land use characteristics within the vicinity of each intersection as well as network characteristics were considered as explanatory variables (Pulugurtha and Nujjetty, 2012).

Wei and Lovegrove (2012) developed Negative Binomial distribution based count models using urban data from the Central Okanagan Regional District (CORD) in Canada. Chiou and Fu (2013) developed a series of Negative Binomial distribution based crash

frequency models to predict for each crash severity level. It should be noted that such an approach can generate interdependence due to latent factors that exist across crash rates at different severity levels (Ma et al., 2008). For example, an increase in one type of severity is also associated with changes in other type of severity.

Wang et al. (2013) studied the effect of access design and spatial pattern on crash risk to pedestrians and bicyclists at access points on urban multilane highways by developing Negative Binomial distribution and logistic regression models to predict crash frequency and injury severity, respectively (Wang et al., 2013).

However, the limitation of the Negative Binomial distribution based model is that time variations are not well considered. Therefore, the standard error of the regression coefficients may be underestimated and the t-ratios may be inflated. Shankar et al. (1997) have attempted to solve this problem by introducing a trend variable using random effects Negative Binomial (RENB) model, which takes into account temporal variability in crash data. In another effort, Ma et al. (2015) analyzed the crash frequency on a freeway using RENB model and explored the effect of various crash contributing factors. The goodness-of-fit statistics showed that RENB model is better than a Negative Binomial distribution based model for the considered dataset.

The presence of many zero crashes in a sample could create problems, which could be tested using the "Vuong non-nested test", in order to develop zero-inflated models (based on Poisson or Negative Binomial distribution) (Pulugurtha and Thakur, 2015).

Overall, Poisson, Negative Binomial and zero-inflated Negative Binomial distribution based models have been applied for analyzing crash frequency data, while

multinomial logit and other discrete choice models have been used to analyze crash injury severities.

2.4 Limitations of Past Research

Several researchers in the past have focused on identifying high bicycle-vehicle crash locations or examining spatial association between bicycle-vehicle crashes and selected factors. Many researchers have investigated the role of various risk factors on bicycle-vehicle crashes and injury risk to bicyclists on roads.

A few researchers have developed safety performance functions or models to estimate crash frequency. However, they focused on developing models for intersections using bicycle counts as an explanatory variable or at TAZ level (Wei and Lovegrove, 2012; Nordback et al., 2014; Nashad et al., 2016). Neither land use data nor detailed network characteristics such as lane miles by speed limit or number of lanes were considered widely in the past studies. Further, the size of the TAZs could vary based on area type (very large in suburban areas) while the characteristics within a TAZ may not be as homogenous as planned.

Undoubtedly, bicycle counts and traffic volume are good predictors of bicycle-vehicle crashes or crashes, in general. However, bicycle counts are yet not typically collected by local agencies. AADT is only available for selected locations (with permanent count stations) in urban areas. Some local agencies collect traffic volume at selected intersections as a part of their data collection programs. However, traffic volume data is not available for most minor arterial streets, collector roads, and local roads. This forces transportation planners and engineers to rely on surrogate data such as demographic, land use and network characteristics to model exposure and/or crash frequency.

The factors that were investigated in the past include demographic, land use and network characteristics; mostly, either individually or in selected combinations. Strong correlations may exist between such characteristics. As an example, the number of lanes (indicator of traffic volume served by a link) may be correlated to the speed limit. However, such correlations were not examined to reduce multicollinearity effect and estimate bicycle-vehicle crash frequency models. This would also limit the data collection efforts while not compromising on the accuracy of estimates. Further, some explanatory variables (example, network characteristics) may be better predictors of bicycle-vehicle crashes than others.

Overall, not many researchers have focused on safety performance functions, i.e., models to estimate bicycle-vehicle crash frequency that would help proactively plan and improve bicyclist safety on urban roads. This dissertation aims to contribute to the body of knowledge by focusing on these aspects.

CHAPTER 3: METHODOLOGY

The proposed research methodology is divided into the following steps.

1. Collect data
2. Identify study locations
3. Generate geospatial buffers around selected locations
4. Extract demographic characteristics within each selected location
5. Extract land use characteristics within each selected location
6. Extract network characteristics within each selected location
7. Develop bicycle-vehicle crash frequency models
8. Validate bicycle-vehicle crash frequency models

The aforementioned steps are discussed next in detail.

3.1 Collect Data

Mecklenburg County, North Carolina was considered as the study area for this research. The study area includes the city of Charlotte as well as towns of Cornelius, Davidson, Huntersville, Matthews, Mint Hill and Pineville.

The data used in this study are obtained from three different sources (all pertaining to local agencies). They include: Charlotte Department of Transportation (CDOT), City of Charlotte Website and Charlotte-Mecklenburg Planning Department.

Crash data for multiple years was used as it would minimize abnormal fluctuation of crashes for a certain year as well as regression to the mean effect often described in safety literature (Emaasit et al., 2013). In this research, bicycle-vehicle crash data from

January 2010 to December 2015 (6 years) was obtained from CDOT. One of the limitations of the crash data obtained from CDOT is the lack of information to identify the age-group of those involved in the crashes.

The demographic, land use and urban road network characteristics were obtained from the Charlotte-Mecklenburg Planning Department and the City of Charlotte website. Transit system characteristics, Charlotte Mecklenburg Schools (CMS) data, and some census data were downloaded from the City of Charlotte Website.

All data was obtained in geospatial format. The data was projected to State Plane Coordinate System; NAD 1983.

3.2 Identify Study Locations

As stated previously, the locations for data extraction, analysis and modeling should be randomly selected to avoid any bias in assessing the relationship, ranking, and allocation of resources for transportation improvements. The number of locations (intersections) must be large enough to yield meaningful outcomes. The following criteria were adopted to ensure that the sample is representative of the study area and its characteristics.

1) The selected locations must be geographically distributed covering all area types (central business district, urban and suburban) in the study area.
2) The selected locations should capture at least 90% of the bicycle-vehicle crashes in the study area.
3) The selected locations must include high risk, medium risk, low risk and no risk locations in the study area. While advanced methods such as network KDE are available, Euclidean KDE method may be adopted to generate crash density map and ensure that the locations are distributed in various risk areas. The method

measures the density of bicycle-vehicle crashes in the vicinity of each reference point over the space. The shading in the map corresponds to the magnitude of Kernel density. Various grid cell sizes and radii must be considered to select values that best capture bicycle-vehicle crash locations.

4) The buffers generated around the selected locations should extract data for most part of the study area.

3.3 Generate Geospatial Buffers around Selected Locations

Geospatial buffers (e.g., 1-mile) are then generated around each identified and selected location in the study area. Multiple buffer widths could be considered to select the best buffer width that can capture geospatial data and help estimate bicycle-vehicle crash frequency. However, 1-mile was identified based on past research as suitable buffer width and considered in this research.

Spatial overlay was then performed to capture geospatial data. Overlay superimposes one map feature over another to create a map feature that has the attributes of both input layers. Clip, intersect, and union are special cases of overlay.

3.4 Extract Demographic Characteristics within the Vicinity of Each Selected Location

Bicycling activity in an area depends on demographic characteristics. The demographic characteristics considered in this study include population, number of household units and household mean-income. To extract demographic characteristics, associated GIS layer with population, household units and household mean-income was overlaid on buffers generated around each selected location. Data extracted was processed as outlined by Pulugurtha and Repaka (2008, 2011).

3.5 Extract Land Use Characteristics within the Vicinity of Each Selected Location

Bicycling activity depends on the land use characteristics as well. The land use characteristics considered include 17 zone classes. They are related to business, business park, business distribution, mixed use, mixed use residential, light industrial, heavy industrial, manufactured home, single-family, multi-family, institutional, research, commercial, office, transit oriented development, uptown mixed use, and urban residential. The land use data was overlaid on the generated buffers to extract the type of land use and type of development in each buffer and examine the role of each selected land use type on bicycle-vehicle crash frequency.

3.6 Extract Network Characteristics within the Vicinity of Each Selected Location

Traffic crashes are most likely to occur at locations with more conflicts. These locations include signalized and unsignalized intersections. To extract these characteristics, layers of the geocoded intersections are overlaid on generated buffers around each selected location so that the spatial joint of above two layers returns a summary of the numeric attributes of the point that fall inside each buffer.

Risk to bicyclists could be higher on high speed and wide roads. Since traffic volume is not available for all the links in the network and past studies showed that traffic volume is related to road design characteristics, they were used as surrogate data. The total length in terms of center-line miles with bicycle lane, without bicycle lane, with sidewalk, without sidewalk, with divided road, with undivided road, with one lane, with two lanes, with three lanes, with 4 lanes, with 5 lanes, with 6 lanes, with 7 lanes, with 25 mph, with 30 mph, with 35 mph, with 40 mph, with 45 mph, with 50 mph, with 55 mph, with 60 mph

and with 65 mph was extracted by overlaying the street centerline network with road design characteristics on the generated buffers.

To access transit system (bus or light-rail), people generally walk, use bicycle or get dropped off. Bus-stops related spatial file was, therefore, overlaid on the generated buffers. Like in the case of intersections, the overlay was used to extract the number of bus-stops in each buffer.

Children and teenagers may walk to schools and colleges. They may have the risk of getting involved in bicycle-vehicle crashes. To extract the number of schools, layers of geocoded schools (elementary, middle, private, high, and college/university) are overlaid on the generated buffers. The number of schools, each type, was extracted and recorded for analysis and modeling.

3.7 Develop Bicycle-Vehicle Crash Frequency Models

Data for the selected locations was divided into modeling dataset and validation dataset to perform analysis, develop statistical models and validate the performance of the models. A Pearson correlation matrix was first developed to examine correlations and select combinations of explanatory variables for modeling. Stepwise eliminations are needed to positively assess Deviance and Pearson Chi-Square ratios in the bicycle-vehicle crash frequency model. Stepwise elimination involves removing statistically insignificant explanatory variables one at a time. This process is repeated until only statistically significant explanatory variables remain in the final model. Models were also developed by considering all explanatory variables irrespective of correlations and significance.

Two types of models (depending on the probability distribution) are considered to establish the relationship between crash frequency and bicyclist's safety risk factors. They are: Poisson and Negative Binomial log-link distribution based models.

3.7.1 Poisson Log-link Distribution Based Model

Considering the discrete, sporadic, non-negative integer character, and random characteristics of crash counts, count-data models such as the Poisson log-link distribution based model appears to be suitable and have been used by many researchers (Chiou et al., 2013; Ma et al., 2015). The Poisson log-link distribution is a discrete probability distribution for the counts of events that occur randomly in a given interval of time or space. In estimating the relative crash frequencies across road sections, it is recommended that the Poisson log-link distribution based model be used as an initial model for developing the relationship (Miaou, 1994).

One requirement of the Poisson log-link distribution is that the mean of the count process equals its variance (Shankar et al., 1995), which is also the limitation of the Poisson log-link distribution based model. Often, many researchers found that the variance is much greater than the mean, indicating that crash data may be over-dispersed. To overcome the over-dispersion problem, researchers have applied the Negative Binomial log-link distribution instead of the Poisson log-link distribution (Miaou, 1994; Shankar et al., 1997; Poch and Mannering, 1996; Abdel-Aty and Radwan, 2000; Ma et al., 2015).

3.7.2 Negative Binomial Log-link Distribution Based Model

The Negative Binomial log-link distribution based model can be used if data are over-dispersed. This model is more efficient than Poisson log-link distribution based model, but in practice the benefits over Poisson are small (Fils, 2012). However, the

Negative Binomial log-link distribution based model should be used if one wishes to predict probabilities and not just model the mean. This model allows the variance to exceed the mean and the Poisson log-link distribution based model can be regarded as a limiting model of the Negative Binomial log-link distribution based model. Although the Negative Binomial log-link distribution based model is more general than the Poisson log-link distribution based model, it requires more extensive computations to estimate model parameters and to generate inferential statistics than the Poisson log-link distribution based model (Miaou, 1994).

3.8 Validate Bicycle-Vehicle Crash Frequency Models

In general, a model is usually developed to analyze a particular problem and used for predictive purposes. A model may represent different parts (assumptions, input parameter values, output values and conclusions) of the system at different levels of abstraction. Performing validation increases the confidence in prediction ability and establishes the credibility of the model. However, in practice, it may be difficult to achieve such a full validation of the model, especially if the system being modelled does not yet exist. Overall, the validation step determines whether the research truly measures what it was intended to measure, or how truthful the research results are.

In this research, mean forecast error (MFE), mean absolute deviation (MAD), mean square error (MSE), root mean square error (RMSE), mean absolute percent error (MAPE) and symmetric mean absolute percent error (SMAPE) are computed to validate the developed models using the validation dataset. They are represented using equations 1 to 5.

$$MAD = \frac{\sum_{i=1}^{n} |A_i - F_i|}{n} \quad \text{… Equation (1)}$$

$$MSE = \frac{\sum_{i=1}^{n}(A_i - F_i)^2}{n} \quad \text{… Equation (2)}$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{n}(A_i - F_i)^2}{n}} \quad \text{… Equation (3)}$$

$$MAPE = \frac{\sum_{i=1}^{n}\left|\frac{A_i - F_i}{A_i}\right|}{n} \times 100 \quad \text{… Equation (4)}$$

$$SMAPE = \frac{\sum_{i=1}^{n}|F_i - A_i|}{\sum_{i=1}^{n}(A_i + F_i)} \quad \text{… Equation (5)}$$

where, A_i is the actual number of bicycle-vehicle crashes in a buffer "i", F_i is the estimated number of crashes in the buffer "i", and n is the number of buffers or study locations for validation.

MAD is the sum of absolute differences between the actual value and the estimated value, divided by the number of observations. MAD indicates by how many units the estimated values differ from the actual values. In general, the smaller the MAD, the better is the model.

MSE, the most used error metric, penalizes larger errors because squaring larger numbers has a greater effect than squaring smaller numbers. It is the sum of the squared errors divided by the number of observations. MSE is typically compared to a standard value, or, between models or methods. The lower the MSE, the better is the model. RMSE is the square root of the MSE.

MAPE is a measure of prediction accuracy of a forecasting method or model. It cannot be used if the actual values includes zeros, because there would be a division by zero. Moreover, MAPE puts a heavier penalty on negative errors (actual value < estimated value).

SMAPE, on the other hand, is an accuracy measure based on percentage (or relative) error. The limitation to SMAPE is that if the actual value or estimated value is zero, the value of error will boom up to the upper-limit of error.

CHAPTER 4: SELECTION OF LOCATIONS AND EXPLANATORY VARIABLES FOR MODELING

As stated previously, all data was obtained in geospatial format and projected to State Plane Coordinate System; NAD 1983. The crash data obtained for Mecklenburg County, North Carolina indicates that there were 628 bicycle-vehicle crashes during the study period. This includes 7 fatal, 15 injury type A, 288 injury type B, 274 injury type C and 44 PDO bicycle-vehicle crashes. Figure 4 shows the spatial distribution of bicycle-vehicle crashes during the study period. This data was used for analysis and modeling. The results obtained from the selection of locations, generation of buffers, data extraction and statistical analysis are presented and discussed in this chapter.

4.1 Selection of Locations

One hundred and nineteen locations (intersections) were selected such that they are geographically distributed throughout the study area (Figure 5). These locations cover 91.8% of the bicycle-vehicle crashes in the study area.

A Kernel density map was generated to overlay and ensure that the selected locations represent high risk, medium risk, low risk and no risk locations. Various cell sizes and radii were tested to generate the Kernel density map. The map based on default cell size (~520.9 feet) and radius (~4,341.0 feet) was considered for illustration and spatial overlay (Figure 6).

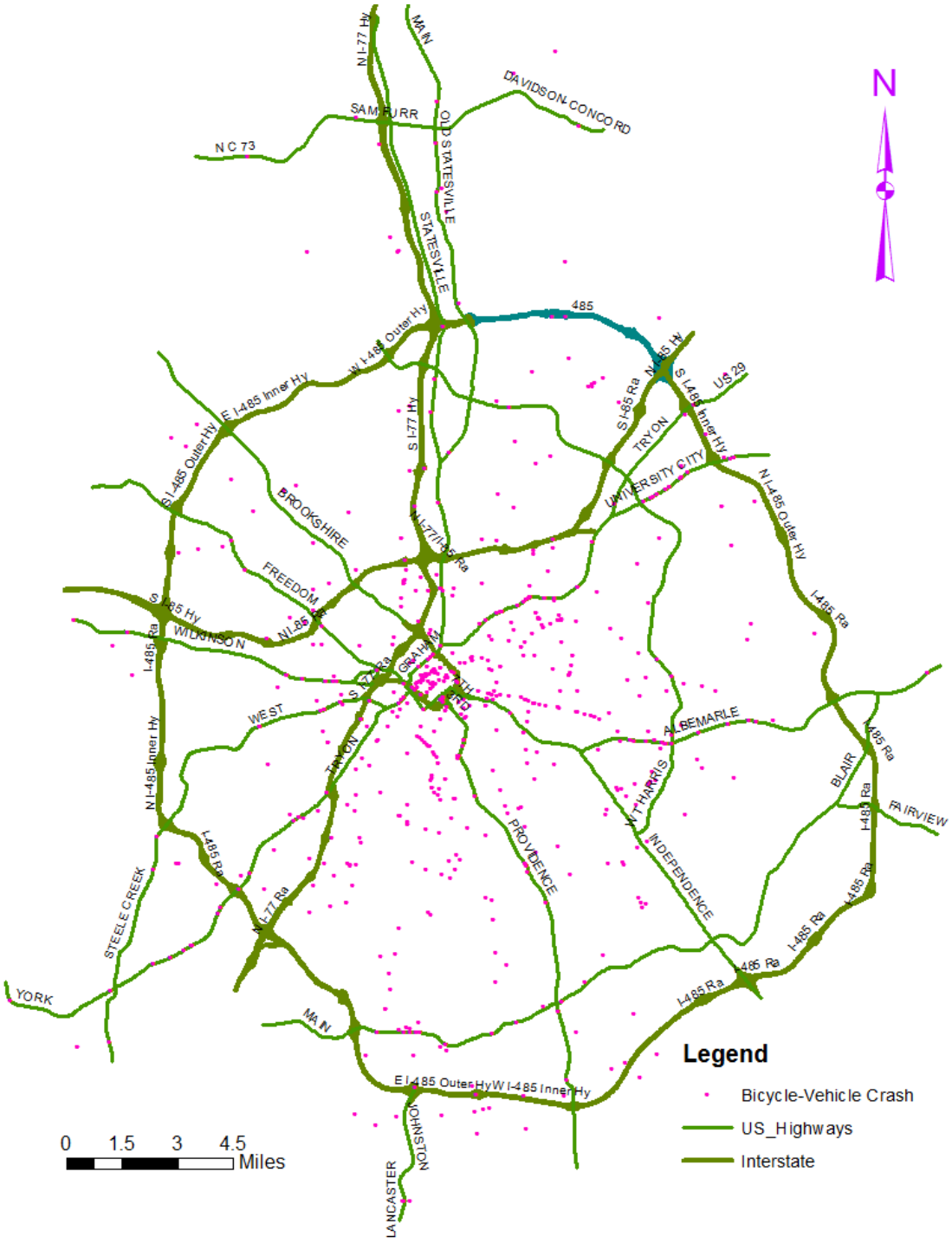

FIGURE 4: Bicycle-vehicle crashes

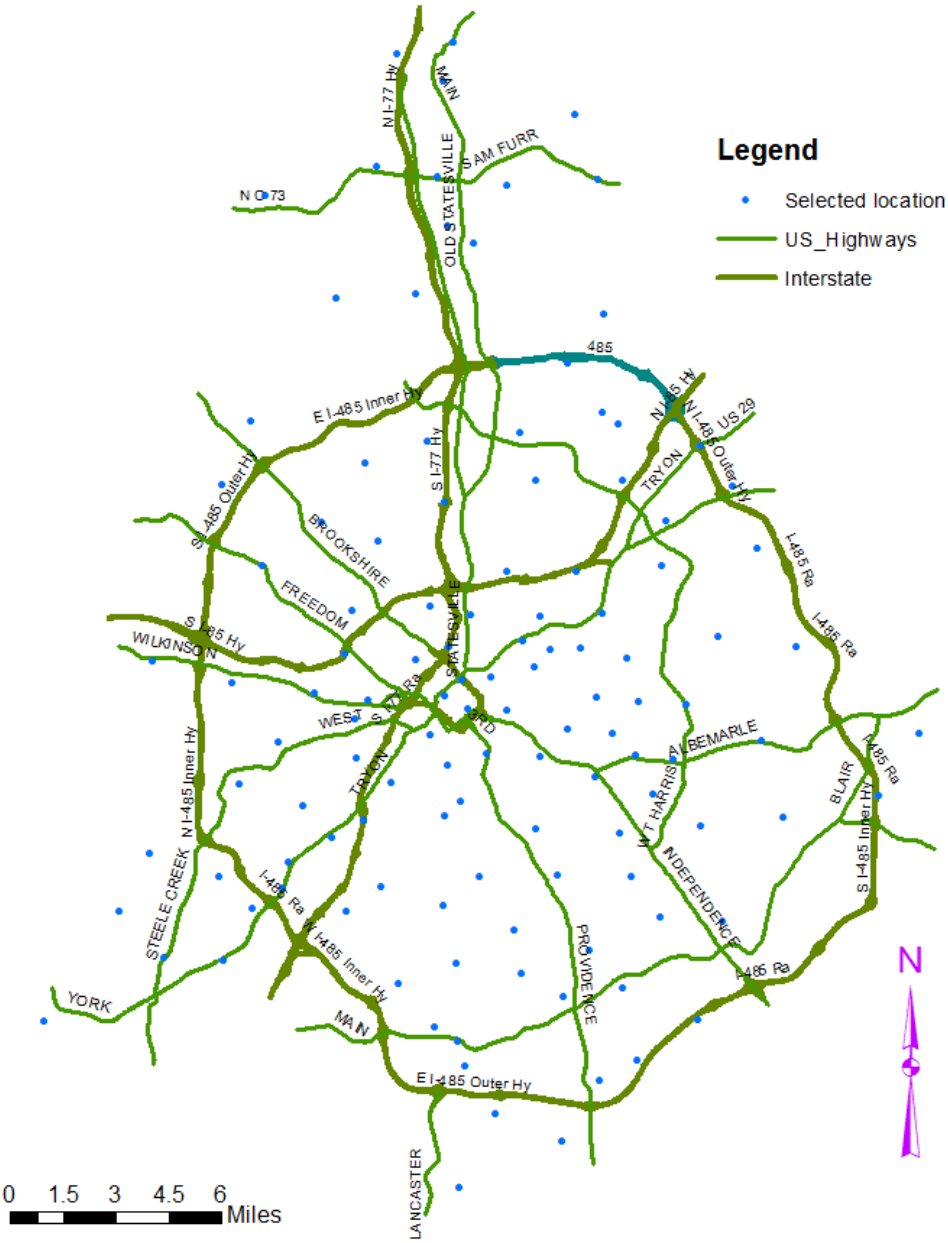

FIGURE 5: Selected study locations

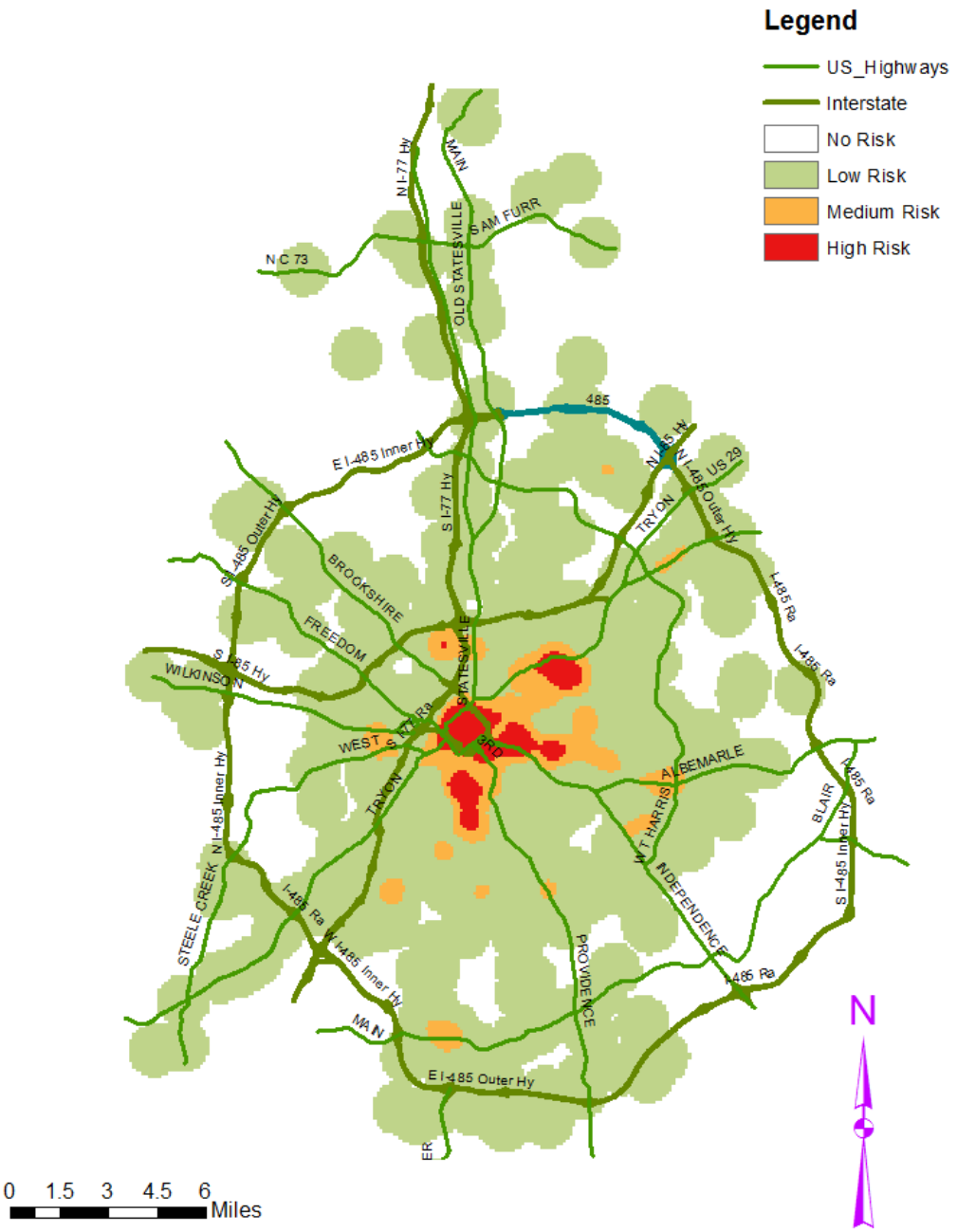

FIGURE 6: Bicycle-vehicle crashes Kernel density

The following symbology was used to define risk in the generated Kernel density map.

No risk	0 to 6 bicycle-vehicle crashes per square mile during the study period
Low risk	6 to 12 bicycle-vehicle crashes per square mile during the study period
Medium risk	12 to 18 bicycle-vehicle crashes per square mile during the study period
High risk	> 18 bicycle vehicle crashes per square mile during the study period

Of the selected 119 locations, 103 locations have seen at least one bicycle-vehicle crash during the study period. Figure 7 depicts spatial overlay of selected locations on bicycle-vehicle crashes and Kernel density.

Buffers of width equal to 1-mile were generated around each selected location (Figure 8). The generated buffers are then spatially overlaid on demographic, land use and network characteristics data to extract data and conduct statistical analysis (figures 9 to 16). The GIS based method presented in Pulugurtha and Repaka (2008) and Pulugurtha and Nujjetty (2012) was adopted in this research to process geospatial data (such as demographic, land use, and network characteristics) and develop databases for analysis and modeling.

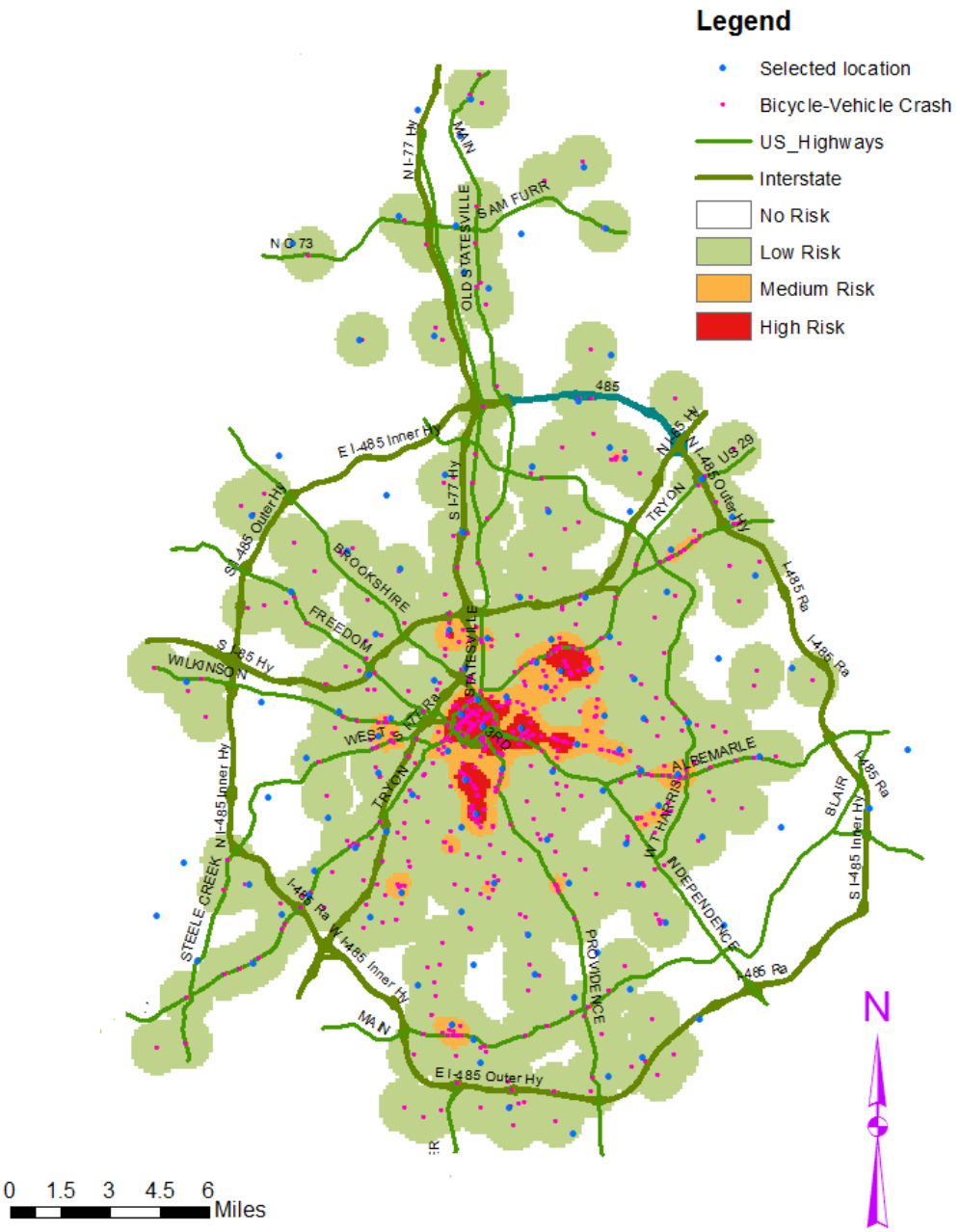

FIGURE 7: Selected locations overlay on bicycle-vehicle crashes and Kernel density

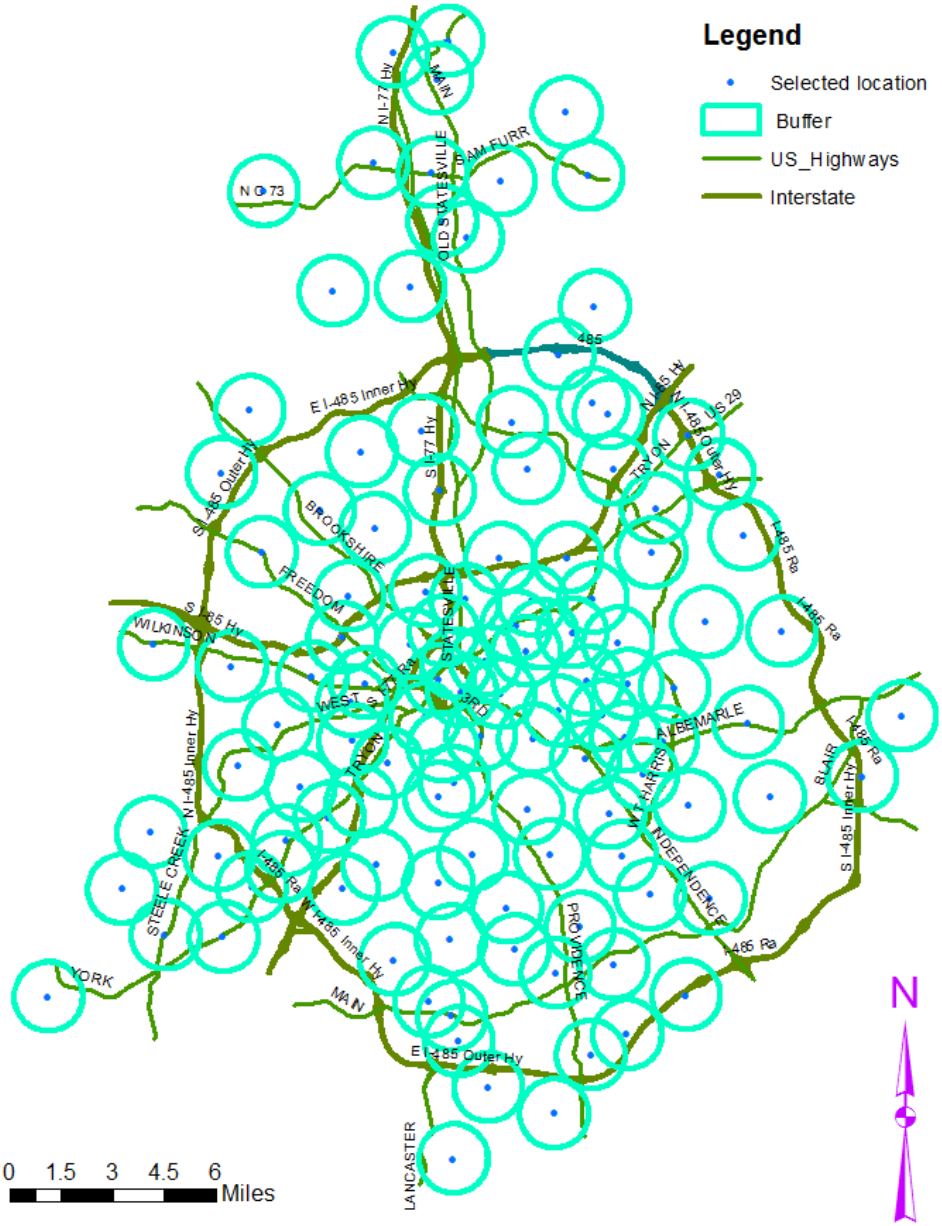

FIGURE 8: Buffers around selected locations

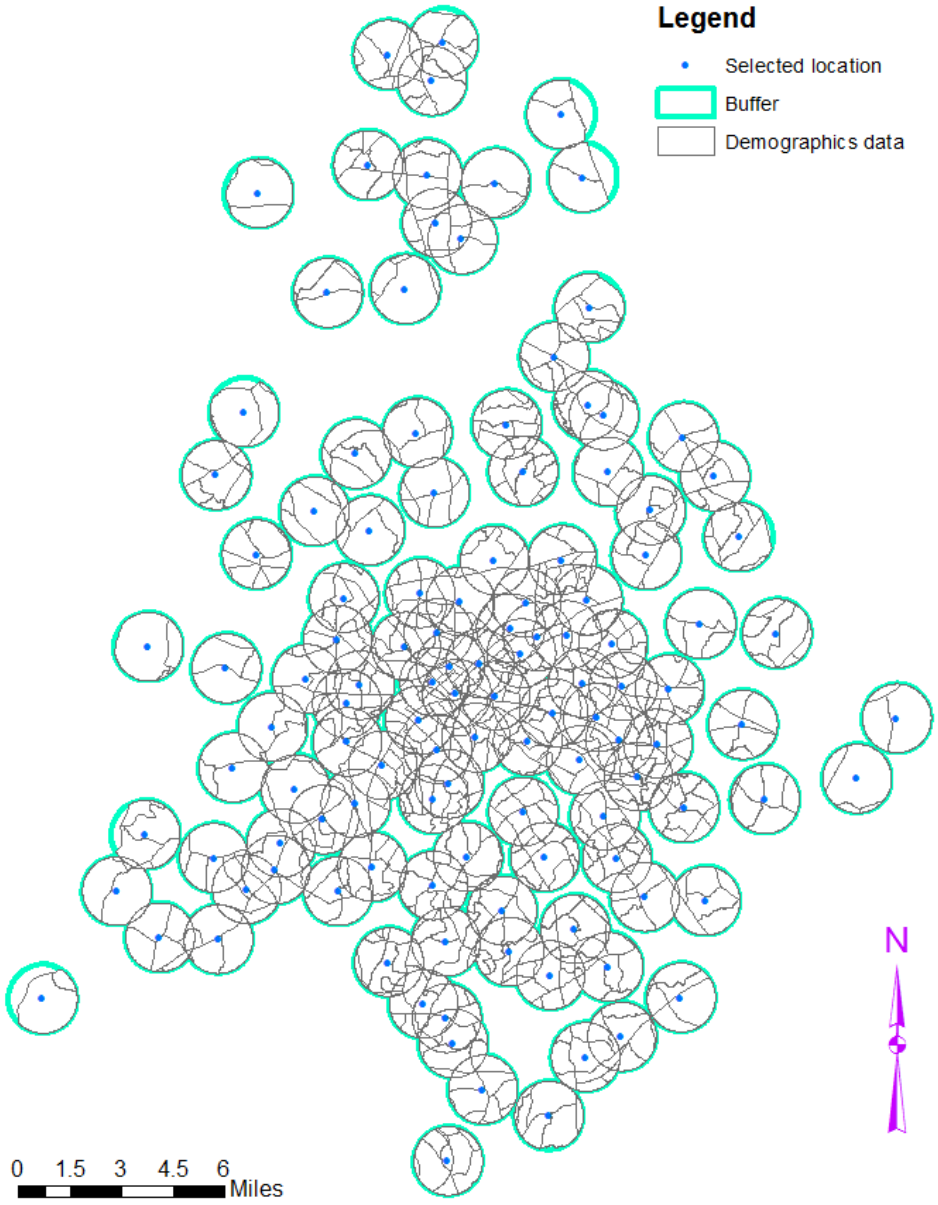

FIGURE 9: Buffers intersected with demographics data

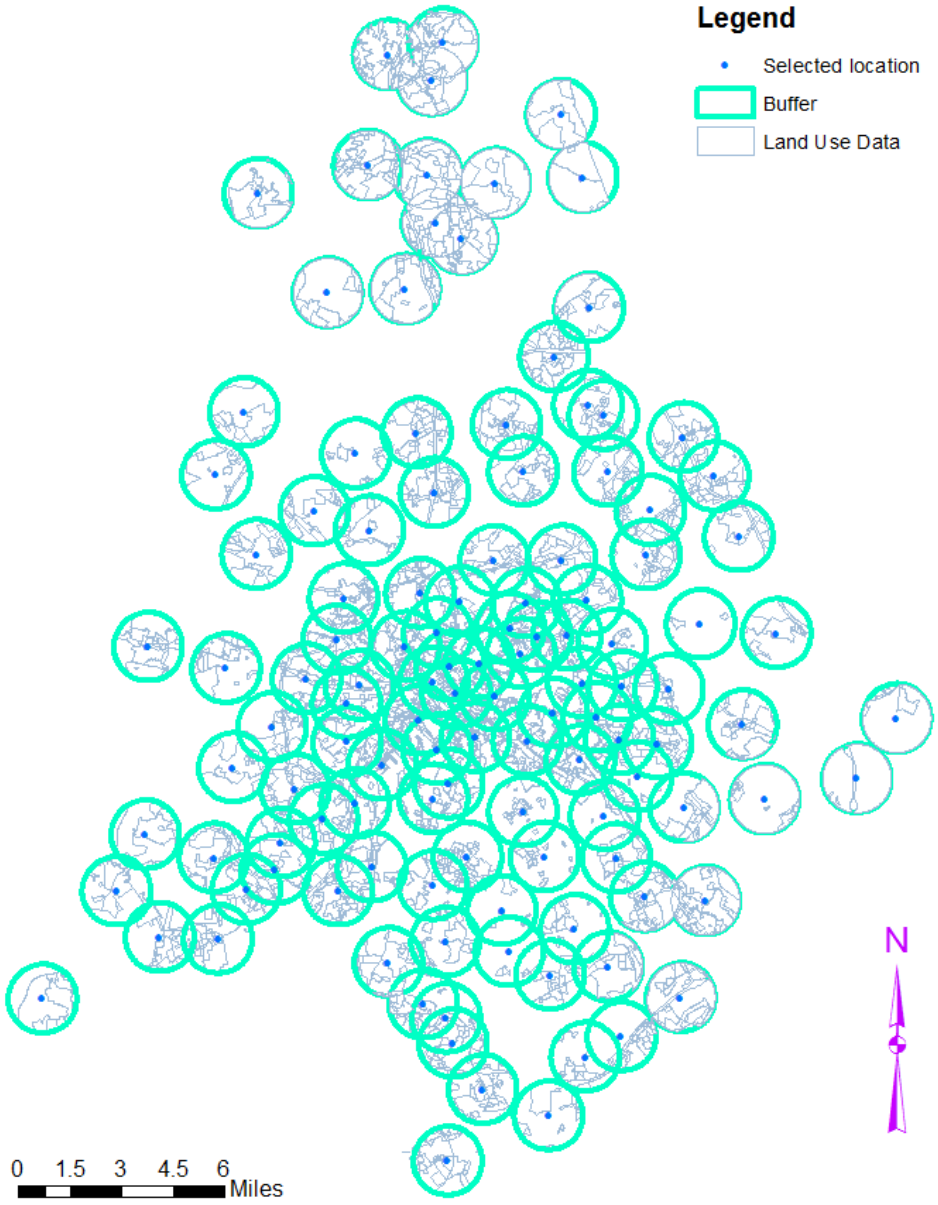

FIGURE 10: Buffers intersected with land use data

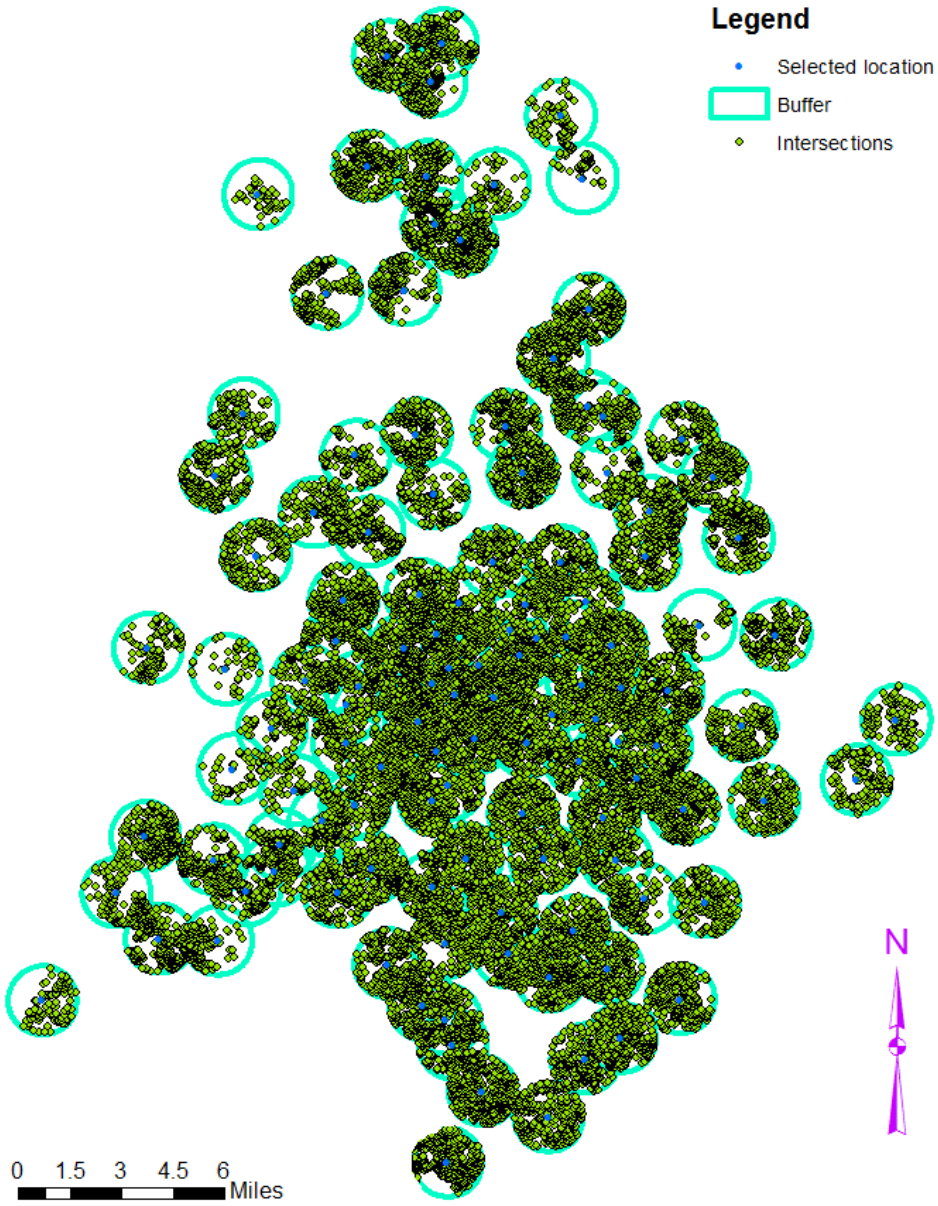

FIGURE 11: Buffers and intersections overlay

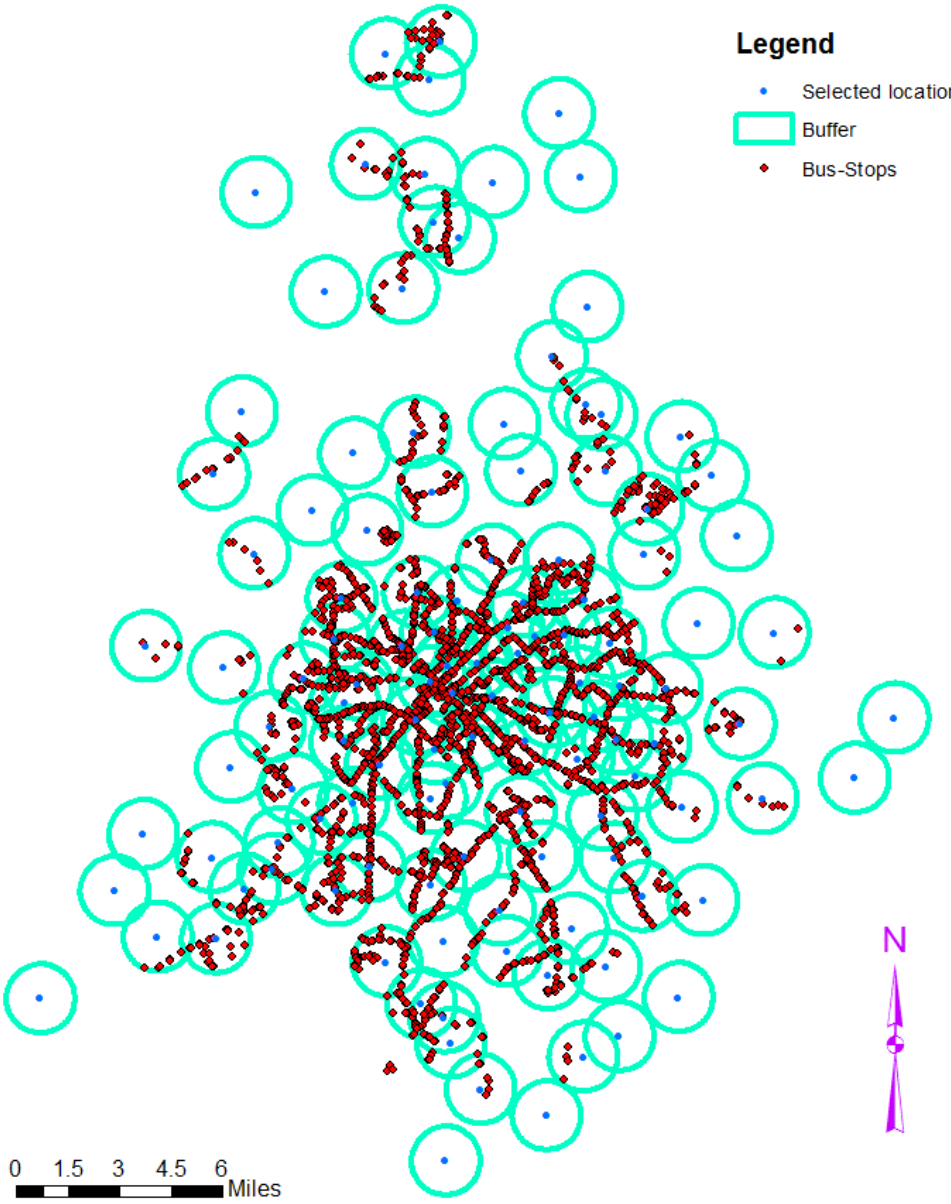

FIGURE 12: Buffers and bus-stops overlay

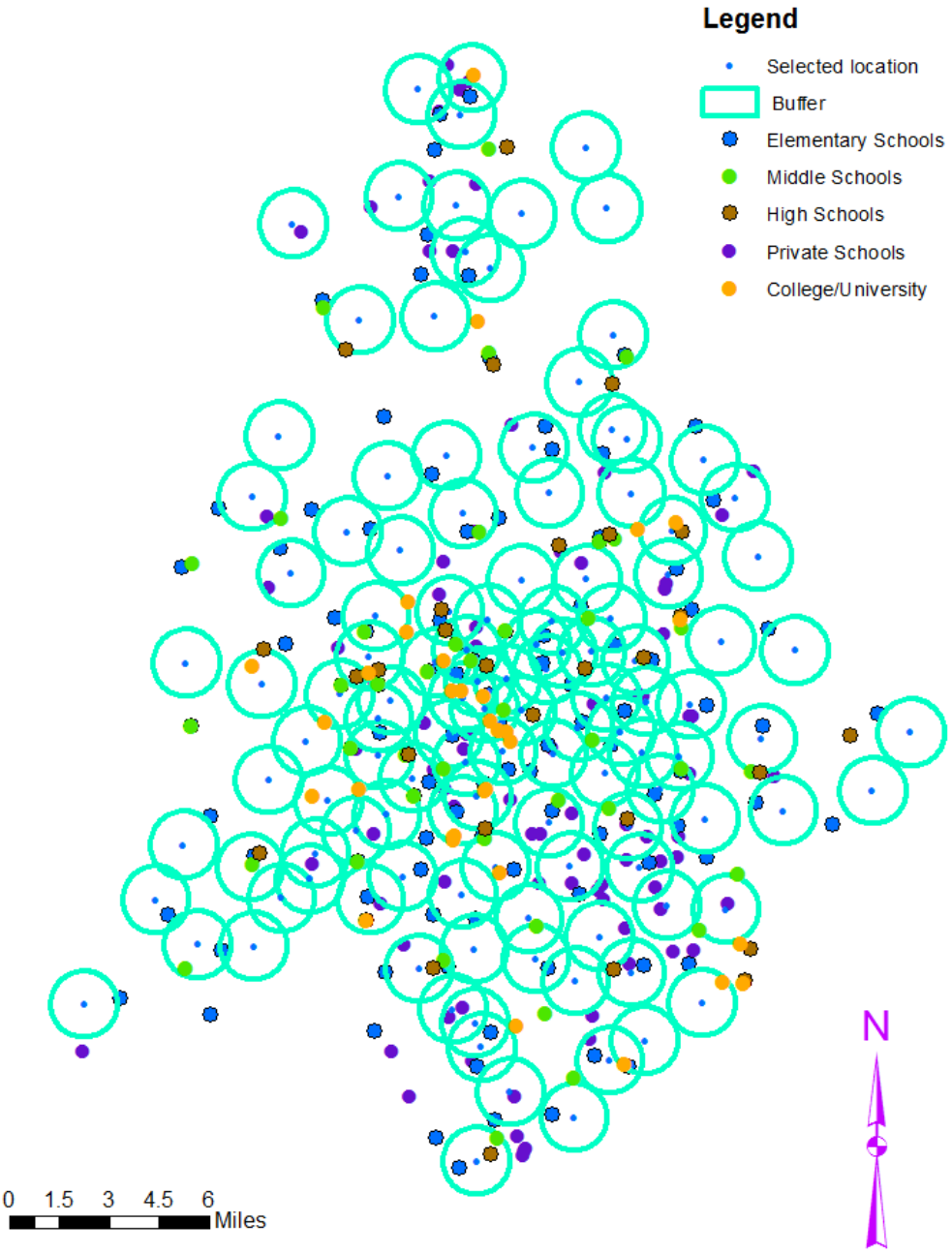

FIGURE 13: Buffers and schools overlay

FIGURE 14: Buffers and streets overlay

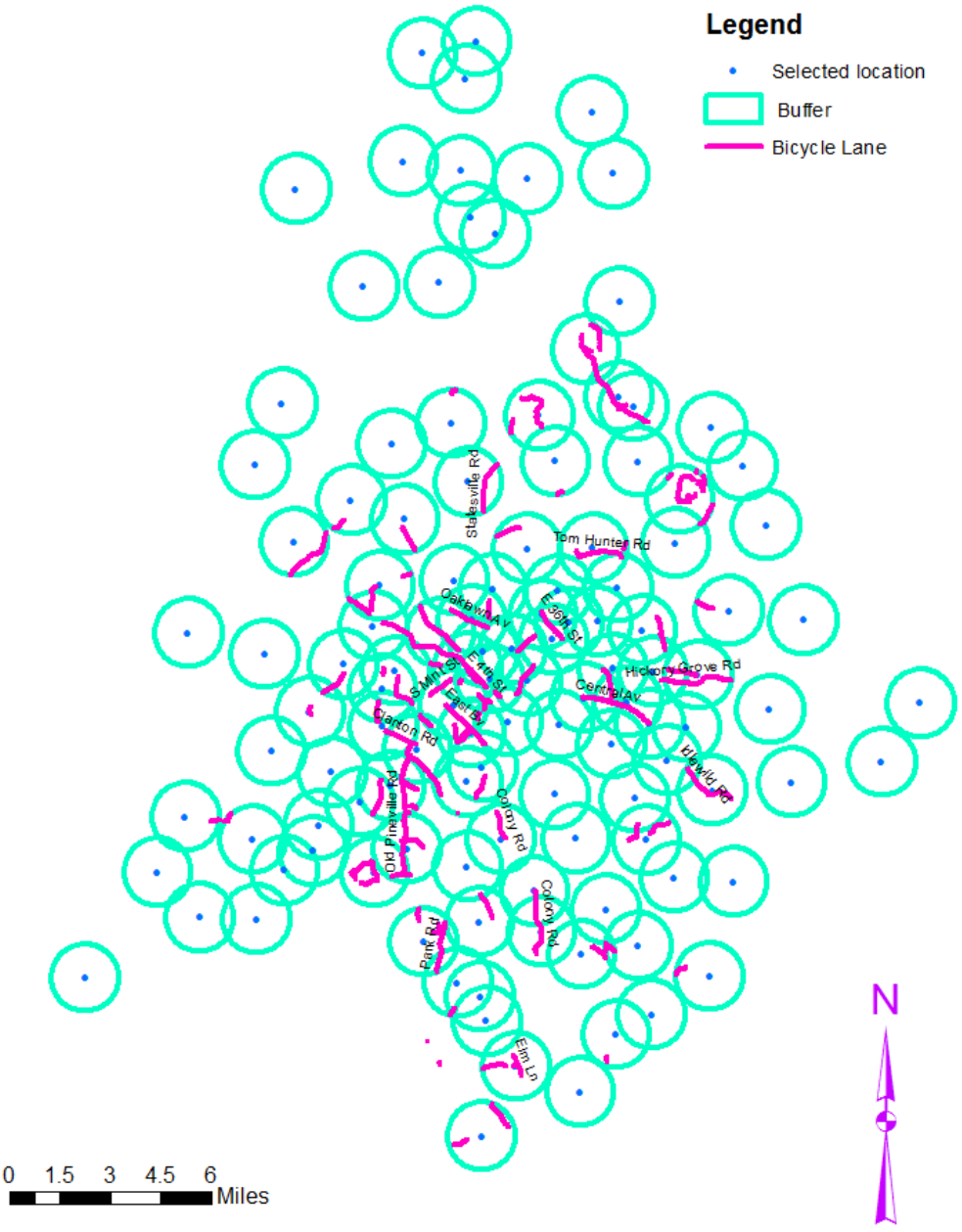

FIGURE 15: Buffers and bicycle / no bicycle lanes overlay

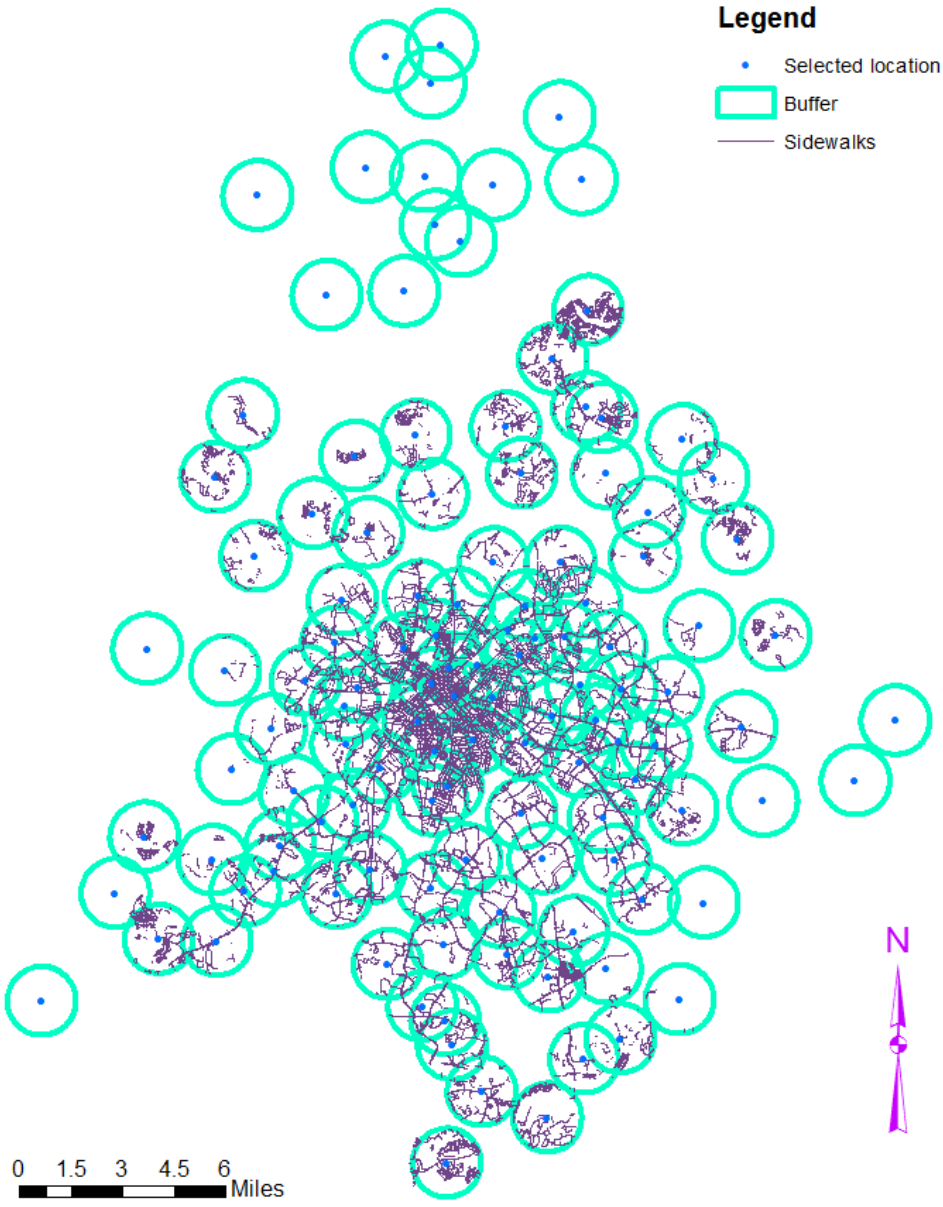

FIGURE 16: Buffers and sidewalks / no sidewalks overlay

4.2 Correlation between Dependent and Explanatory Variables

Several Statistical Software Packages (SAS, STATA, R, and SPSS) may be used to examine correlations between explanatory variables and develop bicycle-vehicle crash frequency models. In this research, IBM SPSS ver. 23 was selected to examine correlations between explanatory variables and develop models. The dependent variable is bicycle-vehicle crash frequency (the number of bicycle-vehicle crashes) during the six-year study period within a 1-mile buffer of the location. A summary of 55 explanatory variables considered in this research along with a brief description is provided in Table 1. As stated previously, the explanatory variables considered include demographic, land use and network characteristics.

Correlations are measures of linear association between two explanatory variables. A correlation test is performed successively among two explanatory variables until all possible combinations has been exhausted. The correlation between two variables may be plotted or graphed into two dimensional spaces under a linear form. However, correlation coefficient is not an appropriate statistic for measuring the association if the relationship is not linear.

Two variables are considered to be strongly correlated to each other if the computed Pearson correlation coefficient is typically less than -0.30 or greater than +0.30. Alternatively, one could consider that there is a significant correlation between two variables if the p-value is less than 0.05 (95% confidence level) or 0.01 (99% confidence level). In this research, an even more conservative approach was adopted. Two explanatory variables were considered to be strongly correlated to each other if the computed p-value is ~0.000 (~100% confidence level).

TABLE 1: Dependent variable and list of explanatory variables

Dependent variable				
NC	Bicycle-vehicle crash frequency (number of bicycle-vehicle crashes)			
Explanatory variables				
Variable	Description		Variable	Description
(a) Network characteristics			(b) Land use characteristics	
IT1	Number of cul-de-sacs		BUS	Area with businesses
IT3	Number of one-way stops on the minor street		BUSPK	Area with business parks
IT4	Number of dead-ends		BUSDIS	Area with business distributions
IT5	Number of traffic lights		MU	Area with mixed use
IT6	Number of road-blocks / private property gates		MUR	Area with mixed use residential
IT7	Number of roundabout loops		LI	Area with light industrial
BS	Number of bus stops		HI	Area with heavy industrial
ES	Number of elementary schools		MH	Area with manufactured home
MS	Number of middle schools		SF	Area with single-family
HS	Number of high schools		MF	Area with multi-family
PS	Number of private schools		INS	Area with institutional
CU	Number of colleges/universities		RES	Area with research
BL	Center-line miles with bicycle lane		COM	Area with commercial
NBL	Center-line miles with no bicycle lane		OFF	Area with office
SW	Center-line miles with sidewalk		TOD	Area with transit oriented
NSW	Center-line miles with no sidewalk		UMU	Area with uptown mixed use
DR	Center-line miles with divided road		UR	Area with urban residential
UDR	Center-line miles with undivided road		(c) Demographic characteristics	
L1	Center-line miles with 1 lane		POP	Total number of population
L2	Center-line miles with 2 lanes		HU	Number of household units
L3	Center-line miles with 3 lanes		MHI	Mean household income
L4	Center-line miles with 4 lanes			
L5	Center-line miles with 5 lanes			
L6	Center-line miles with 6 lanes			
L7	Center-line miles with 7 lanes			
L8	Center-line miles with 8 lanes			
25mph	Center-line miles with 25 mph as speed limit			
30mph	Center-line miles with 30 mph as speed limit			
35mph	Center-line miles with 35 mph as speed limit			
40mph	Center-line miles with 40 mph as speed limit			
45mph	Center-line miles with 45 mph as speed limit			
50mph	Center-line miles with 50 mph as speed limit			
55mph	Center-line miles with 55 mph as speed limit			
60mph	Center-line miles with 60 mph as speed limit			
65mph	Center-line miles with 65 mph as speed limit			

Table 2 summarizes Pearson correlation coefficients computed between explanatory variables as well as between the dependent variable and explanatory variables considered in this research. Shaded cells in the table indicate that p-value is ~0.000 (strong correlation).

The bicycle-vehicle crash frequency is linearly correlated to all considered network characteristics except the number of dead-ends (IT4) and center-line miles with speed limit 45 mph, 60 mph or 65 mph at a 95% or higher confidence level. It is linearly correlated to business (BUS), mixed use residential (MUR), heavy industrial (HI), single-family (SF), multi-family (MF), office (OFF), transit oriented development (TOD), uptown mixed use (UMU) and urban residential (UR) areas at a 95% or higher confidence level. All considered demographic characteristics are linearly correlated to bicycle-vehicle crash frequency at a 95% or higher confidence level. With exceptions of the number of cul-de-sacs (IT1), mixed use residential (MUR) area and single-family (SF) residential area, an increase in all other explanatory variables could lead to an increase in bicycle-vehicle crash frequency.

TABLE 2: Pearson correlation coefficients – Summary

a) Network characteristics

Variable	NC	IT1	IT3	IT4	IT5	IT6	IT7	BS	ES	MS	HS	PS	CU	BL	NBL	SW	NSW	DR	UDR	L1	L2	L3	L4	L5	L6	L7	L8	25mph	30mph	35mph	40mph	45mph	50mph	55mph	60mph	65mph
NC	1	-.216*	.610**	0.042	.814**	.292**	.491**	.817**	.538**	.299**	.389**	.350**	.558**	.462**	.644**	.760**	-0.137	.414**	.640**	.402**	.484**	.578**	.748**	.563**	.259**	.216*	.521**	.623**	.716**	.365**	.217*	0.123	.245**	.437**	0.161	-0.097
IT1		1	0.096	0.066	-0.098	0.128	-0.091	-.341**	-0.087	-0.086	-0.074	0.008	-0.065	-0.063	0.069	-0.070	.269**	-0.051	0.076	-0.167	.246**	-0.068	-.255**	-0.150	0.007	0.026	-0.040	0.001	-0.084	0.172	.188*	0.066	0.100	-.236**	-0.058	.323**
IT3			1	.226*	.495**	.320**	.389**	.667**	.582**	.376**	.377**	.391**	.305**	.461**	.932**	.729**	.212*	.414**	.946**	.317**	.931**	.375**	.562**	.424**	0.168	0.156	.315**	.682**	.438**	.742**	.317**	0.067	0.059	.317**	0.156	-0.076
IT4				1	.208*	-0.070	.425**	0.099	-0.023	.186*	.293**	0.122	.222*	0.093	.231*	0.054	.336**	0.072	.237**	.422**	.207*	.185*	0.162	-0.030	0.115	0.114	0.168	-0.129	0.126	.467**	0.010	0.060	.194*	.281**	.242**	.253**
IT5					1	.282**	.734**	.787**	.470**	.394**	.431**	.382**	.636**	.558**	.580**	.655**	0.016	.501**	.562**	.623**	.362**	.790**	.780**	.546**	.336**	.393**	.531**	.485**	.804**	.408**	.196*	0.129	.335**	.637**	.252**	0.033
IT6						1	0.127	.342**	.411**	.343**	.267**	0.106	0.135	.236**	.342**	.350**	0.051	.244**	.332**	0.120	.341**	.178*	.300**	0.115	.403**	0.153	.202*	.358**	.255**	0.166	.381**	.244**	0.174	0.136	.310**	-0.003
IT7							1	.548**	.308**	.458**	.464**	.335**	.453**	.463**	.465**	.460**	0.112	.322**	.471**	.790**	.274**	.715**	.547**	.288**	.203*	.183*	.372**	.230*	.593**	.443**	0.083	-0.031	.316**	.742**	.238**	.199*
BS								1	.649**	.528**	.509**	.389**	.504**	.581**	.733**	.778**	-0.052	.597**	.706**	.566**	.525**	.636**	.844**	.580**	.385**	.253**	.517**	.671**	.709**	.404**	.186*	.184*	.245**	.619**	.313**	-.186*
ES									1	.568**	.397**	.288**	.343**	.491**	.593**	.634**	-0.069	.361**	.602**	.303**	.520**	.410**	.533**	.329**	.216*	0.142	.280**	.579**	.527**	.308**	.241**	0.108	0.126	.435**	0.121	-0.165
MS										1	.624**	.221*	.243**	.491**	.428**	.497**	-0.027	.394**	.419**	.535**	.308**	.477**	.471**	.263**	.327**	0.170	.297**	.401**	.374**	.238**	0.138	0.161	.184*	.588**	.317**	-0.101
HS											1	.222*	.400**	.387**	.410**	.467**	-0.025	.352**	.397**	.501**	.289**	.435**	.459**	.193*	.312**	0.078	.253**	.298**	.381**	.305**	0.106	0.099	.209*	.531**	.396**	-0.075
PS												1	.297**	.234**	.427**	.284**	.256**	.351**	.398**	.261**	.353**	.324**	.455**	.301**	.201*	0.110	.279**	.407**	.303**	.274**	0.078	0.044	.215*	.270**	.211*	0.017
CU													1	.292**	.377**	.455**	-0.006	.302**	.362**	.364**	.269**	.460**	.480**	.380**	.211*	.216*	.466**	.393**	.506**	.239**	0.042	-0.010	.302**	.412**	.213*	0.057
BL														1	.446**	.588**	-0.067	.466**	.480**	.449**	.385**	.592**	.503**	.386**	.203*	.203*	.275**	.526**	.570**	.227*	.355**	0.117	.179*	.552**	0.157	-0.140
NBL															1	.757**	.248**	.561**	.982**	.466**	.925**	.460**	.698**	.461**	.215*	0.159	.328**	.707**	.501**	.744**	.240**	0.104	0.099	.417**	.268**	0.033
SW																1	-.393**	.440**	.769**	.413**	.624**	.564**	.708**	.546**	.287**	.250**	.459**	.819**	.561**	.308**	.350**	0.170	0.158	.520**	0.148	-.194*
NSW																	1	.184*	.211*	0.154	.343**	-0.035	0.000	-0.099	0.004	0.013	-0.047	-.193*	0.039	.585**	-0.070	-0.036	0.112	-0.057	.287**	.408**
DR																		1	.420**	.518**	.380**	.430**	.675**	.464**	.367**	.277**	.285**	.367**	.406**	.352**	0.164	.347**	0.139	.519**	.373**	0.062
UDR																			1	.426**	.937**	.455**	.643**	.426**	0.166	0.127	.311**	.725**	.501**	.732**	.257**	0.043	0.090	.386**	.214*	0.003
L1																				1	.229*	.603**	.647**	.262**	0.140	0.148	.368**	.180*	.542**	.405**	0.040	0.090	.197*	.831**	.387**	.224*
L2																					1	.227*	.438**	.300**	0.116	0.089	.238**	.663**	.332**	.744**	.287**	0.054	0.057	.180*	.192*	0.061
L3																						1	.578**	.372**	.281**	.230*	.402**	.367**	.752**	.298**	0.137	0.055	.374**	.699**	.183*	0.118
L4																							1	.569**	.294**	.259**	.417**	.572**	.648**	.410**	0.174	.241**	.197*	.644**	.428**	-0.127
L5																								1	.290**	.515**	.337**	.554**	.383**	0.165	.190*	.230*	0.119	.290**	0.086	-0.153
L6																									1	.470**	.218*	.285**	.226*	0.048	.284**	.330**	.428**	.191*	.242**	-0.003
L7																										1	.306**	.262**	.226*	0.061	.300**	0.130	.227*	0.161	0.137	0.089
L8																											1	.389**	.462**	.179*	.214*	0.041	.358**	.421**	.202*	0.047
25mph																												1	.414**	0.129	.335**	0.038	0.166	.260**	0.137	-.183*
30mph																													1	.333**	.225*	0.002	.380**	.626**	.207*	0.027
35mph																														1	0.055	-0.037	0.022	.288**	.231*	0.121
40mph																															1	0.126	0.109	0.054	0.027	0.056
45mph																																1	0.143	0.088	0.174	0.006
50mph																																	1	.221*	.275**	.229*
55mph																																		1	.284**	-0.065
60mph																																			1	0.048
65mph																																				1

*. Correlation is significant at the 0.05 level (2-tailed).

**. Correlation is significant at the 0.01 level (2-tailed).

TABLE 2: Pearson correlation coefficients – Summary (continued)

b) Land use characteristics

Variable	NC	BUS	BUSPK	BUSDIS	MU	MUR	LI	HI	MH	SF	MF	INS	RES	COM	OFF	TOD	UMU	UR
NC	1	.430**	-0.104	-0.085	0.051	-.271**	-0.066	.239**	-0.127	-.226*	.308**	0.006	-0.075	-0.099	.321**	.229*	.476**	.588**
BUS		1	-0.002	0.032	-0.089	-.315**	0.035	0.107	0.098	-.197*	.349**	0.046	0.017	-0.101	.283**	.257**	0.035	0.139
BUSPK			1	.473**	0.016	-0.112	0.123	-0.071	-0.063	-0.032	0.010	0.074	-0.038	0.075	0.171	-0.020	-0.049	-0.089
BUSDIS				1	0.017	-0.116	0.137	0.103	-0.047	-0.168	0.039	0.101	-0.012	-0.044	0.174	0.138	-0.049	-0.072
MU					1	0.105	0.009	-0.049	-0.068	-.349**	-0.150	-0.047	0.006	0.005	0.042	0.026	0.156	0.035
MUR						1	-0.083	-.237**	-0.004	-0.174	-.331**	-0.111	-0.055	.273**	-.233*	-0.111	-0.002	-0.125
LI							1	.296**	0.076	-.458**	-0.030	-0.040	-0.075	0.171	-0.091	-0.001	-0.064	-0.035
HI								1	-0.011	-.482**	-0.011	-0.073	-0.082	-.205*	-0.172	0.154	0.135	.290**
MH									1	0.054	-.194*	-0.033	-0.033	-0.052	-0.110	-0.018	-0.043	-0.080
SF										1	-.216*	-0.164	-0.172	-.345**	-0.091	-0.178	-.236**	-.246**
MF											1	0.162	-0.049	0.105	.380**	0.085	-0.098	0.030
INS												1	0.023	.281**	0.074	0.001	-0.042	-0.017
RES													1	.246**	0.055	-0.011	-0.026	-0.051
COM														1	0.172	-0.068	-0.105	-0.135
OFF															1	0.101	0.002	-0.039
TOD																1	.255**	-0.006
UMU																	1	.610**
UR																		1

**. Correlation is significant at the 0.01 level (2-tailed).

*. Correlation is significant at the 0.05 level (2-tailed).

c) Demographic characteristics

Variable	NC	POP	HU	MHI
NC	1	.453**	.690**	-.469**
POP		1	.392**	-.740**
HU			1	-.456**
MHI				1

4.3 Selection of Explanatory Variables for Modeling

Many strong correlations were observed between the selected network characteristics. Business area (BUS) was observed to be strongly correlated to mixed use residential (MUR) and multi-family (MF) areas. Single-family (SF) area was observed to be strongly correlated to mixed use (MU), light industrial (LI), heavy industrial (HI), and commercial (COM) areas. Multi-family area (MF) was observed to be strongly correlated to business (BUS), mixed use residential (MUR) and office (OFF) areas. Business park area (BUSPK) was observed to be strongly correlated to business district area (BUSDIS), while uptown mixed use area (UMU) was observed to be strongly correlated to urban residential area (UR). Household units (HU) and mean household income (MHI) were observed to be strongly correlated to the total population (POP).

To minimize multicollinearity, only one explanatory variable of two explanatory variables that are strongly correlated was considered for modeling. Several combinations could be built when selecting / eliminating explanatory variables for modeling. This will help generate alternate models and test their predictability as well as validity.

The number of traffic lights (IT5) and the number of one-way stops on the minor street (IT3) are critical variables but strongly correlated to each other. Therefore, the number of traffic lights (IT5) was considered in models 3, 4, and 5, while the number of one-way stops on the minor street (IT3) was considered in model 6, 7 and 8. Network characteristics that are not strongly correlated to the selected critical variable were considered in the respective models.

Land use variables that are not strongly correlated to mixed use (MU) and multi-family (MF) areas were considered in models 3 and 6, while land use variables that are not

strongly correlated to single-family (SF) and multi-family (MF) areas were considered in models 4 and 7. Land use variables that are not strongly correlated to mixed use (MU) and office (OFF) areas were considered in models 5 and 8.

The number of miles without bicycle lane (NBL) and the number of miles with no sidewalk (NSW) were considered in all the models.

In addition to the combinations of explanatory variables, two other models were also developed ignoring the strong correlations between explanatory variables. Except in case of Model 1, statistically insignificant variables were eliminated one after another in all others until the model has only significant explanatory variables.

Table 3 summarizes the combinations of explanatory variables that were considered for modeling in this research.

TABLE 3: Summary of selected explanatory variable combinations for modeling

Explanatory Variable	Models 1 & 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8
IT1	X	X	X	X	X	X	X
IT3	X				X	X	X
IT4	X	X	X	X	X	X	X
IT5	X	X	X	X			
IT6	X	X	X	X			
IT7	X						
BS	X						
ES	X						
MS	X						
HS	X						
PS	X						
CU	X				X	X	X
BL	X	X	X	X	X	X	X
NBL	X	X	X	X	X	X	X
SW	X						
NSW	X	X	X	X	X	X	X
DR	X						
UDR	X						
L1	X						
L2	X						
L3	X						
L4	X						
L5	X						
L6	X				X	X	X
L7	X				X	X	X
L8	X						
25mph	X						
30mph	X						
35mph	X						
40mph	X	X	X	X			
45mph	X	X	X	X	X	X	X
50mph	X				X	X	X
55mph	X						
60mph	X	X	X	X	X	X	X
65mph	X	X	X	X	X	X	X
BUS	X			X			X
BUSPK	X						
BUSDIS	X	X	X	X	X	X	X
MU	X	X			X		
MUR	X						
LI	X	X			X		
HI	X	X			X		
MH	X	X	X	X	X	X	X
SF	X		X	X		X	X
MF	X	X	X		X	X	
INS	X	X	X	X	X	X	X
RES	X	X	X	X	X	X	X
COM	X	X			X		
OFF	X			X			X
TOD	X	X	X	X	X	X	X
UMU	X	X			X		
UR	X		X	X		X	X
POP	X	X	X	X	X	X	X
HU	X						
MHI	X						

CHAPTER 5: BICYCLE-VEHICLE CRASH FREQUENCY MODELS

The results obtained from the development of bicycle-vehicle crash frequency models are discussed in this chapter.

In this research, 119 locations were randomly selected to cover 91.8% of bicycle-vehicle crashes in the study area. Of these, 99 randomly selected locations were used to develop bicycle-vehicle crash frequency models. The remaining 20 randomly selected locations were used to validate the predictability of the developed bicycle-vehicle crash frequency models.

5.1 Test the Applicability of Poisson Log-link Distribution for Model 1

Based on literature review, many researchers revealed that Poisson distribution is suitable and most frequently applied to model crash frequency (Ma et al., 2015; Wang et al., 2015; Anastasopoulos et al., 2008; and Ivan et al., 2000). To paraphrase Miaou's (1994) comment, "In estimating the relative crash frequencies across road sections, it is recommended that the Poisson regression model be used as an initial model for developing the relationship." The limitation of the Poisson model is that the mean must be equal to the variance (Miaou, 1994; Shankar et al., 1995; Vogt and Bared, 1998).

Firstly, tests were conducted to investigate if Poisson distribution would best fit the data. SPSS software recommends two tests: (1) "One-Sample Kolomogorov-Smirnov Test" (K-S Test), which is a non-parametric test; and, (2) a "Descriptive Statistics: Mean and Variance" test. These two tests reveal if the observed count (crash) data follows the Poisson distribution.

Table 4 summarizes results obtained from “One Sample Kolmogorov-Smirnov Test” and “Descriptive Statistics” test. The computed p-value from the “One Sample Kolmogorov-Smirnov Test” is ~0.000, indicating that the bicycle-vehicle crash data does not follow a Poisson distribution. From the “Descriptive Statistics” test, the mean (6.56) is much smaller than the variance (56.52). The computed dispersion parameter is 1.139, indicating that the bicycle-vehicle crash data is over-dispersed. Such a relation was observed by several researchers in the past (Miaou, 1994; Shankar et al., 1995; Vogt and Bared, 1998). As stated previously, to overcome the over-dispersion problem, researchers have applied the Negative Binomial log-link distribution instead of the Poisson log-link distribution (Miaou, 1994; Shankar et al., 1997; Poch and Mannering, 1996; Abdel-Aty and Radwan, 2000; Ma et al., 2015).

A model using Poisson log-link distribution was developed even though the bicycle-vehicle crash data was observed to be over-dispersed. Table 5 summarizes the Poisson regression coefficients along with their standard errors (Std. Error), Wald Chi-Square values, p-values (significance values) and goodness-of-fit statistics (Deviance, Pearson Chi-Square, Akaike’s Information Criterion (AIC) and Finite Sample Corrected AIC (AICC)) for the Poisson log-link distribution based bicycle-vehicle crash frequency Model 1. All explanatory variables were considered irrespective of possible correlations between themselves. Further, the explanatory variables were not eliminated even if they have a statistically insignificant effect on bicycle-vehicle crash frequency.

TABLE 4: Testing applicability of Poisson log-link distribution for modeling

NPar Tests

[DataSet0] \\filer.uncc.edu\home\kkmukoko\BicycleDataset15.sav

Descriptive Statistics

	N	Mean	Std. Deviation	Minimum	Maximum
NC	97	6.5567	7.51799	.00	32.00

One-Sample Kolmogorov-Smirnov Test

		NC
N		97
Normal Parameters[a,b]	Mean	6.5567
	Std. Deviation	7.51799
Most Extreme Differences	Absolute	.229
	Positive	.229
	Negative	-.192
Test Statistic		.229
Asymp. Sig. (2-tailed)		.000[c]

a. Test distribution is Normal.

b. Calculated from data.

c. Lilliefors Significance Correction.

Descriptives

Descriptive Statistics

	N	Mean	Variance
NC	97	6.5567	56.520
Valid N (listwise)	97		

TABLE 5: Poisson log-link distribution based Model 1 - results summary

Parameter	Coeff.	Std. Error	Wald Chi-Square	p-value	D (value/df)	PCS (value/df)	AIC	AICC
(Intercept)	-0.603	2.29	0.07	0.79	1.9	1.7	496.4	648.4
IT1	-0.003	0.00	0.58	0.45				
IT3	0.01	0.00	6.27	0.01				
IT4	-0.012	0.01	2.51	0.11				
IT5	0.005	0.02	0.07	0.79				
IT6	0.037	0.05	0.48	0.49				
IT7	-0.031	0.03	1.44	0.23				
BS	-0.008	0.01	1.70	0.19				
ES	0.192	0.12	2.62	0.11				
MS	-0.157	0.15	1.13	0.29				
HS	0.169	0.19	0.83	0.36				
PS	0.201	0.14	1.94	0.16				
CU	0.075	0.10	0.60	0.44				
BL	-17.83	14.50	1.51	0.22				
NBL	-17.593	14.46	1.48	0.22				
SW	19.144	14.21	1.82	0.18				
NSW	19.114	14.20	1.81	0.18				
DR	0.016	0.09	0.03	0.86				
UDR								
L1	-0.268	0.72	0.14	0.71				
L2	-0.602	0.71	0.73	0.39				
L3	-0.571	0.70	0.67	0.42				
L4	-0.629	0.73	0.74	0.39				
L5	-0.03	0.73	0.00	0.97				
L6	-0.29	0.72	0.16	0.69				
L7	-1.974	0.83	5.68	0.02				
L8	-1.141	0.72	2.52	0.11				
25mph	-0.93	0.45	4.20	0.04				
30mph	-0.613	0.48	1.65	0.20				
35mph	-0.968	0.45	4.54	0.03				
40mph	-1.165	0.48	5.95	0.02				
45mph	-0.924	0.46	4.00	0.05				
50mph	-1.204	0.67	3.24	0.07				
55mph	-1.112	0.44	6.36	0.01				

TABLE 5: Poisson log-link distribution based Model 1 - summary (continued)

60mph	-0.981	0.44	4.98	0.03				
65mph	-0.834	0.43	3.80	0.05				
BUS	9.86E-08	0.00	5.05	0.03				
BUSPK	5.06E-08	0.00	0.19	0.67				
BUSDIS	-1.15E-07	0.00	2.47	0.12				
MU	1.10E-09	0.00	0.00	0.98				
MUR	-1.33E-08	0.00	0.22	0.64				
LI	7.90E-09	0.00	0.10	0.76				
HI	2.19E-08	0.00	0.75	0.39				
MH	1.95E-08	0.00	0.06	0.81				
SF	-3.59E-09	0.00	0.02	0.89				
MF	1.61E-08	0.00	0.22	0.64				
INS	6.13E-08	0.00	4.33	0.04				
RES	1.00E-07	0.00	9.20	<0.01				
COM	5.37E-08	0.00	0.80	0.37				
OFF	-9.96E-09	0.00	0.08	0.77				
TOD	-9.89E-08	0.00	0.81	0.37				
UMU	3.02E-08	0.00	0.12	0.73				
UR	-4.09E-08	0.00	0.11	0.74				
POP	5.44E-06	0.00	0.03	0.87				
HU	-6.20E-05	0.00	4.27	0.04				
MHI	8.67E-06	0.00	0.34	0.56				

Note: D = Deviance, PCS = Pearson Chi-Square, AIC = Akaike's Information Criterion, AICC = Finite Sample Corrected AIC

The Omnibus Test indicated that the developed Poisson distribution based Model 1 is statistically significant. The coefficients for one-way stop on the minor street (IT3), area with businesses (BUS), area with institutional (INS) and area with research (RES) are positive and significant at a 95% confidence level. This means that for each one-unit increase of above explanatory variables, the expected log count of bicycle-vehicle crash frequency (NC) increases by its respective coefficient (e.g., 0.010 for IT3). The coefficients for miles with 7 lanes (L7), miles with 25 mph as speed limit (25mph), miles

with 35 mph as speed limit (35mph), miles with 40 mph as speed limit (40mph), miles with 45 mph as speed limit (45mph), miles with 55 mph as speed limit (55mph), miles with 60 mph as speed limit (60mph), miles with 65 mph as speed limit (65mph) and household units (HU) are negative and significant at a 95% confidence level. For each one-unit increase of above explanatory variables, the expected log count of bicycle-vehicle crash frequency (NC) decreases by its respective coefficient. The aforementioned statistically significant explanatory variables are highlighted in Table 5.

The standard errors indicate how much the variable prediction is "off". The smaller the standard errors, the better will be the prediction model. For example, the highest standard error in this Poisson distribution based model is for miles with 7 lanes (L7), for which the prediction is off by "0.83". This is followed by miles with 40 mph as speed limit (40 mph), for which the prediction is off by "0.48". The standard error for one-way stop on the minor street (IT3), area with businesses (BUS), area with institutional (INS), area with research (RES) and household units (HU) is zero (least).

The computed Deviance and Pearson Chi-Square values per degrees of freedom are 1.9 and 1.7, respectively (outside the expected range of 0.9 to 1.1). The computed AIC = 496.4 and AICC = 648.4. The difference between AIC and AICC is very high, indicating that the model does not fit the data well.

5.2 Negative Binomial Log-link Distribution Based Model 1

As test results from both "One-Sample Kolmogorov-Smirnov Test" and "Descriptive Statistics" test indicate that the considered bicycle-vehicle crash data is over-dispersed, a model using Negative Binomial log-link distribution was developed using bicycle-vehicle crash frequency as the dependent variable.

The results obtained for the Negative Binomial log-link distribution based Model 1 are summarized and shown in Table 6. The table shows coefficients for each selected explanatory variable, standard errors (Std. Error), Wald Chi-Square values, p-values (significance values) and goodness-of-fit statistics (Deviance, Pearson Chi-Square, AIC and AICC)) for the Negative Binomial log-link distribution based bicycle-vehicle crash frequency Model 1. All explanatory variables were considered irrespective of possible correlations between themselves. Further, the explanatory variables were not eliminated even if they have a statistically insignificant effect on bicycle-vehicle crash frequency.

The Omnibus Test indicated that the developed Negative Binomial log-link distribution based Model 1 is statistically significant. The coefficients for one-way stop on the minor street (IT3), area with businesses (BUS), area with institutional (INS) and area with research (RES) are positive and significant at a 95% confidence level. This means that for each one-unit increase of above explanatory variables, the expected log count of bicycle-vehicle crash frequency (NC) increases by its respective coefficient (e.g., 0.010 for IT3). The coefficients for miles with 7 lanes (L7), miles with 25 mph as speed limit (25mph), miles with 35 mph as speed limit (35mph), miles with 40 mph as speed limit (40mph), miles with 45 mph as speed limit (45mph), miles with 55 mph as speed limit (55mph), miles with 60 mph as speed limit (60mph), miles with 65 mph as speed limit (65mph) and household units (HU) are negative and significant at a 95% confidence level. For each one-unit increase of above explanatory variables, the expected log count of bicycle-vehicle crash frequency (NC) decreases by its respective coefficient. The aforementioned statistically significant explanatory variables are highlighted in Table 6.

TABLE 6: Negative Binomial log-link distribution based Model 1 - results summary

Parameter	Coeff.	Std. Error	Wald Chi-Square	p-value	D (value/df)	PCS (value/df)	AIC	AICC
(Intercept)	-0.603	2.29	0.07	0.79	1.9	1.8	498.4	659.7
IT1	-0.003	0.00	0.58	0.45				
IT3	0.01	0.00	6.27	0.01				
IT4	-0.012	0.01	2.51	0.11				
IT5	0.005	0.02	0.07	0.79				
IT6	0.037	0.05	0.48	0.49				
IT7	-0.031	0.03	1.44	0.23				
BS	-0.008	0.01	1.70	0.19				
ES	0.192	0.12	2.62	0.11				
MS	-0.157	0.15	1.13	0.29				
HS	0.169	0.19	0.83	0.36				
PS	0.201	0.14	1.94	0.16				
CU	0.075	0.10	0.60	0.44				
BL	-17.83	14.50	1.51	0.22				
NBL	-17.593	14.46	1.48	0.22				
SW	19.144	14.21	1.82	0.18				
NSW	19.114	14.20	1.81	0.18				
DR	0.016	0.09	0.03	0.86				
UDR								
L1	-0.268	0.72	0.14	0.71				
L2	-0.602	0.71	0.73	0.39				
L3	-0.571	0.70	0.67	0.42				
L4	-0.629	0.73	0.74	0.39				
L5	-0.03	0.73	0.00	0.97				
L6	-0.29	0.72	0.16	0.69				
L7	-1.974	0.83	5.68	0.02				
L8	-1.141	0.72	2.52	0.11				
25mph	-0.93	0.45	4.20	0.04				
30mph	-0.613	0.48	1.65	0.20				
35mph	-0.968	0.45	4.54	0.03				
40mph	-1.165	0.48	5.95	0.02				
45mph	-0.924	0.46	4.00	0.05				
50mph	-1.204	0.67	3.24	0.07				
55mph	-1.112	0.44	6.36	0.01				

TABLE 6: Negative Binomial log-link distribution based Model 1 - summary (continued)

60mph	-0.981	0.44	4.98	0.03				
65mph	-0.834	0.43	3.80	0.05				
BUS	9.86E-08	0.00	5.05	0.03				
BUSPK	5.06E-08	0.00	0.19	0.67				
BUSDIS	-1.15E-07	0.00	2.47	0.12				
MU	1.10E-09	0.00	0.00	0.98				
MUR	-1.33E-08	0.00	0.22	0.64				
LI	7.90E-09	0.00	0.10	0.76				
HI	2.19E-08	0.00	0.75	0.39				
MH	1.95E-08	0.00	0.06	0.81				
SF	-3.59E-09	0.00	0.02	0.89				
MF	1.61E-08	0.00	0.22	0.64				
INS	6.13E-08	0.00	4.33	0.04				
RES	1.00E-07	0.00	9.20	<0.01				
COM	5.37E-08	0.00	0.80	0.37				
OFF	-9.96E-09	0.00	0.08	0.77				
TOD	-9.89E-08	0.00	0.81	0.37				
UMU	3.02E-08	0.00	0.12	0.73				
UR	-4.09E-08	0.00	0.11	0.74				
POP	5.44E-06	0.00	0.03	0.87				
HU	-6.20E-05	0.00	4.27	0.04				
MHI	8.67E-06	0.00	0.34	0.56				

Note: D = Deviance, PCS = Pearson Chi-Square, AIC = Akaike's Information Criterion
AICC = Finite Sample Corrected AIC

The highest standard error in this Negative Binomial distribution based model is for miles with 7 lanes (L7), for which the prediction is off by "0.83". This is followed by miles with 40 mph as speed limit (40 mph), for which the prediction is off by "0.48". The standard error for one-way stop on the minor street (IT3), area with businesses (BUS), area with institutional (INS), area with research (RES) and household units (HU) is zero (least).

The computed Deviance and Pearson Chi-Square values per degrees of freedom are 1.9 and 1.8, respectively (outside the expected range of 0.9 to 1.1 as in the case of Poisson

log-link distribution based Model 1). The computed AIC = 498.4 and AICC = 659.7. The difference between AIC and AICC is very high, indicating that the model does not fit the data well. This could be due to the presence of several insignificant explanatory variables or possible correlations between the variables.

The computed dispersion parameter for the Negative Binomial log-link distribution based Model 1 is equal to 1.312E-8. It is very low (almost equal to 0), indicating that using Negative Binomial log-link distribution corrected the over-dispersion problem.

The developed Negative Binomial log-link distribution based Model 1 is expressed in its exponential form as Equation (6).

NC/year = [EXP((-0.603) + (-0.003*IT1)+(0.01*IT3) + (-0.012*IT4) + (0.005*IT5) + (0.037*IT6) + (-0.031*IT7) + (-0.008*BS) + (0.192*ES) + (-0.157*MS) + (0.169*HS) + (0.201*PS) + (0.075*CU) + (-17.83*BL) + (-17.593*NBL) + (19.144*SW) + (19.114*NSW) + (0.016*DR) + (-0.268*L1) + (-0.602*L2) + (-0.571*L3) + (-0.629*L4) + (-0.03*L5) + (-0.29*L6) + (-1.974*L7) + (-1.141*L8) + (-0.93*25mph) + (-0.613*30mph) + (-0.968*35mph) + (-1.165*40mph) + (-0.924*45mph) + (-1.204*50mph) + (-1.112*55mph) + (-0.981*60mph) + (-0.834*65mph) + (0.0000000986*BUS) + (0.0000000506*BUSPK) + (-0.000000115*BUSDIS) + (0.0000000011*MU) + (-0.0000000133*MUR) + (0.0000000079*LI) + (0.0000000219*HI) + (0.0000000195*MH) + (-0.00000000359*SF) + (0.0000000161*MF) + (0.0000000613*INS) + (0.0000001*RES) + (0.0000000537*COM) + (-0.00000000996*OFF) + (-0.0000000989*TOD) + (0.0000000302*UMU) + (-0.0000000409*UR) + (0.00000544*POP) + (-0.000062*HU) + (0.00000867*MHI))]/6 … Equation (6)

Based on interpretation of model parameters and considering conscientious violations for Poisson log-link distribution applicability, this research considers Negative Binomial log-link distribution based models as best models to estimate bicycle-vehicle crash frequency. Only Negative Binomial log-link distribution based models are therefore discussed hereafter.

5.3 Negative Binomial Log-link Distribution Based Model 2

All explanatory variables are considered initially when developing Model 2. However, statistically insignificant explanatory variables are eliminated one after another until the model comprises of only significant explanatory variables ($p<=0.05$, at a 95% confidence level). The results obtained for the developed final Negative Binomial log-link distribution based Model 2 are summarized and shown in Table 7. The table shows coefficients for each selected explanatory variable (unstandardized coefficient), standard errors (Std. Error), Wald Chi-Square values, p-values (significance values) and goodness-of-fit statistics (Deviance, Pearson Chi-Square, AIC and AICC).

TABLE 7: Negative Binomial log-link distribution based Model 2 - results summary

Parameter	Coeff.	Std. Error	Wald Chi-Square	p-value	D (value/df)	PCS (value/df)	AIC	AICC
(Intercept)	-0.851	0.32	7.10	0.01	1.3	1.2	475.9	480.9
IT5	0.034	0.01	35.89	<0.01				
IT7	-0.045	0.01	26.63	<0.01				
NBL	0.089	0.01	76.06	<0.01				
NSW	-0.024	0.01	7.61	0.01				
L7	-1.114	0.39	8.20	<0.01				
L8	-0.478	0.17	8.06	0.01				
35mph	-0.042	0.02	7.46	0.01				
BUS	5.96E-08	0.00	8.34	<0.01				
HI	1.58E-08	0.00	9.23	<0.01				
MF	3.60E-08	0.00	10.73	<0.01				
RES	5.33E-08	0.00	5.16	0.02				
COM	8.97E-08	0.00	10.74	<0.01				

Note: D = Deviance, PCS = Pearson Chi-Square, AIC = Akaike's Information Criterion
AICC = Finite Sample Corrected AIC

The Omnibus Test indicated that the developed Negative Binomial log-link distribution based Model 2 is statistically significant and the model degrees of freedom is twelve (12) explanatory variables. The coefficients for traffic lights (IT5), miles with no bicycle lane (NBL), area with businesses (BUS), area with heavy industrial (HI), area with multi-family (MF), area with research (RES) and area with commercial (COM) are positive. The other explanatory variables such as roundabout loop (IT7), miles with no sidewalk (NSW), miles with 7 lanes (L7), miles with 8 lanes (L8) and miles with 35 mph as speed limit (35mph) have negative coefficients.

The highest standard error in this Negative Binomial log-link distribution based Model 2 is for miles with 7 lanes (L7), for which the prediction is off by "0.3890". This is followed by the intercept, for which the prediction is off by "0.3196". The least standard error is for heavy industrial (HI) area, for which the prediction is off by "5.1999E-9".

The computed Deviance and Pearson Chi-Square values per degrees of freedom are 1.3 and 1.2, respectively (outside the expected range of 0.9 to 1.1 but lower than for Negative Binomial log-link distribution based Model 1). The computed AIC = 475.9 and AICC = 480.9. They are lower than AIC and AICC for Negative Binomial log-link distribution based Model 1. The estimated dispersion parameter for this model is equal to 0.069.

The developed Negative Binomial log-link distribution based Model 2 is expressed in its exponential form as Equation (7).

NC/year = [EXP ((-0.851) + (0.034*IT5) + (-0.045*IT7) + (0.089*NBL) + (-0.024*NSW) + (-1.114*L7) + (-0.478*L8) + (-.042*35mph) + (0.00000005960*BUS) + (0.00000001580*HI) + (0.00000003600*MF) + (0.00000005333*RES) + (0.00000008967*COM))]/6 … Equation (7)

5.4 Negative Binomial Log-Link Distribution Based Model 3

Network characteristics that are not correlated to traffic lights (IT5), land use characteristics that are not correlated to mixed use (MU) and multi-family (MF) areas, and population are considered initially. Statistically insignificant explanatory variables are eliminated one after another until the model comprises of only significant explanatory variables ($p<=0.05$, at a 95% confidence level). The results obtained for the developed final Negative Binomial log-link distribution based Model 3 are summarized and shown in Table 8. The table shows coefficients for each selected explanatory variable (unstandardized coefficient), standard errors (Std. Error), Wald Chi-Square values, p-values (significance values) and goodness-of-fit statistics (Deviance, Pearson Chi-Square, AIC and AICC).

The Omnibus Test indicated that the developed Negative Binomial log-link distribution based Model 3 is statistically significant and the model degrees of freedom is ten (10) explanatory variables. The coefficients for traffic lights (IT5), miles with no bicycle lane (NBL), area with heavy industrial (HI), area with multi-family (MF), area with research (RES) and area with commercial (COM) are positive. The other explanatory variables such as cul-de-sacs (IT1), dead-ends (IT4), miles with no sidewalk (NSW) and area with uptown mixed used (UMU) have negative coefficients.

The highest standard error in this Negative Binomial log-link distribution based model is for the intercept, for which the prediction is off by “0.3764”. This is followed by

traffic lights (IT5), for which the prediction is off by "0.0147". The least standard error is for heavy industrial (HI) area, for which the prediction is off by "6.1276E-9".

The computed Deviance and Pearson Chi-Square values per degrees of freedom are 1.2 and 1.1, respectively (close to the expected range). The computed AIC = 491.8 and AICC = 495.4 for this model. The estimated dispersion parameter for this model is equal to 0.146.

The developed Negative Binomial log-link distribution based Model 3 is expressed in its exponential form as Equation (8).

$$NC/year = [(EXP ((-0.441) + (-0.005*IT1) + (-0.013*IT4) + (0.035*IT5) + (0.062*NBL) + (-0.025*NSW) + (0.00000001325*HI) + (0.00000004101*MF) + (0.00000006220*RES) + (0.00000007675*COM) + (-0.0000001181*UMU))]/6 \quad \ldots \text{Equation (8)}$$

TABLE 8: Negative Binomial log-link distribution based Model 3 - results summary

Parameter	Coeff.	Std. Error	Wald Chi-Square	p-value	D (value/df)	PCS (value/df)	AIC	AICC
(Intercept)	-0.441	0.38	1.37	0.24	1.2	1.1	491.8	495.4
IT1	-0.005	0.00	3.88	0.05				
IT4	-0.013	0.01	6.32	0.01				
IT5	0.035	0.01	5.60	0.02				
NBL	0.062	0.01	37.10	<0.01				
NSW	-0.025	0.01	6.27	0.01				
HI	1.33E-08	0.00	4.68	0.03				
MF	4.10E-08	0.00	11.56	<0.01				
RES	6.22E-08	0.00	5.33	0.02				
COM	7.68E-08	0.00	5.89	0.02				
UMU	-1.18E-07	0.00	4.51	0.03				

Note: D = Deviance, PCS = Pearson Chi-Square, AIC = Akaike's Information Criterion
AICC = Finite Sample Corrected AIC

5.5 Negative Binomial Log-link Distribution Based Model 4

Network characteristics that are not correlated to traffic lights (IT5), land use characteristics that are not correlated to single-family (SF) and multi-family (MF) areas, and population are considered initially. Statistically insignificant explanatory variables are eliminated one after another until the model comprises of only significant explanatory variables ($p<=0.05$, at a 95% confidence level). The results obtained for the developed final Negative Binomial log-link distribution based Model 4 are summarized and shown in Table 9. The table shows contains coefficients for each selected explanatory variable (unstandardized coefficient), standard errors (Std. Error), Wald Chi-Square values, p-values (significance values) and goodness-of-fit statistics (Deviance, Pearson Chi-Square, AIC and AICC).

The Omnibus Test indicated that the developed Negative Binomial log-link distribution based Model 4 is statistically significant and the model degrees of freedom is six (6) explanatory variables.

TABLE 9: Negative Binomial log-link distribution based Model 4 - results summary

Parameter	Coeff.	Std. Error	Wald Chi-Square	p-value	D (value/df)	PCS (value/df)	AIC	AICC
(Intercept)	0.132	0.40	0.11	0.74				
IT4	-0.018	0.01	9.02	<0.01				
NBL	0.081	0.01	131.98	<0.01				
NSW	-0.041	0.01	22.65	<0.01	1.2	1.0	501.4	503.0
SF	-1.19E-08	0.00	10.09	<0.01				
MF	3.68E-08	0.00	8.63	<0.01				
INS	3.17E-08	0.00	4.48	0.03				

Note: D = Deviance, PCS = Pearson Chi-Square, AIC = Akaike's Information Criterion
AICC = Finite Sample Corrected AIC

The coefficients for miles with no bicycle lane (NBL), area with multi-family (MF) and area with institutional (INS) are positive. The other explanatory variables such as dead-ends (IT4), area with no sidewalk (NSW) and area with single-family (SF) have negative coefficients.

The highest standard error in this Negative Binomial log-link distribution based Model 4 is for the intercept, for which the prediction is off by "0.4022". This is followed by the area with no sidewalk (NSW), for which the prediction is off by "0.0085". The least standard error is for single-family (SF) area, for which the prediction is off by "3.7566E-9".

The computed Deviance and Pearson Chi-Square values per degrees of freedom are 1.2 and 1.0, respectively (close to the expected range). The computed AIC = 501.4 and AICC = 503.0 for this model. The estimated dispersion parameter for this model is equal to 0.210.

The developed Negative Binomial log-link distribution based Model 4 is expressed in its exponential form as Equation (9).

$$NC/year = [EXP ((0.132) + (-0.018*IT4) + (0.081*NBL) + (-0.041*NSW) + (-0.00000001193*SF) + (0.00000003678*MF) + (0.00000003173*INS))]/6 \quad \ldots \text{Equation (9)}$$

5.6 Negative Binomial Log-link Distribution Based Model 5

Network characteristics that are not correlated to traffic lights (IT5), land use characteristics that are not correlated to mixed use (MU) and office (OFF) areas, and population are considered initially. Statistically insignificant explanatory variables are eliminated one after another until the model comprises of only significant explanatory

variables (p<=0.05, at a 95% confidence level). The results obtained for the developed final Negative Binomial log-link distribution based Model 5 are summarized and shown in Table 10. The table shows coefficients for each selected explanatory variable (unstandardized coefficient), standard errors (Std. Error), Wald Chi-Square values, p-values (significance values) and goodness-of-fit statistics (Deviance, Pearson Chi-Square, AIC and AICC).

TABLE 10: Negative Binomial log-link distribution based Model 5 - results summary

Parameter	Coeff.	Std. Error	Wald Chi-Square	p-value	D (value/df)	PCS (value/df)	AIC	AICC
(Intercept)	0.159	0.43	0.13	0.72	1.2	1.0	503.5	505.5
IT1	-0.006	0.00	6.40	0.01				
IT4	-0.015	0.01	6.70	0.01				
NBL	0.07	0.01	58.32	<0.01				
NSW	-0.033	0.01	13.72	<0.01				
45mph	0.125	0.05	7.10	0.01				
POP	5.70E-05	0.00	6.33	0.01				
SF	-1.08E-08	0.00	6.35	0.01				

Note: D = Deviance, PCS = Pearson Chi-Square, AIC = Akaike's Information Criterion
AICC = Finite Sample Corrected AIC

The Omnibus Test indicated that the developed Negative Binomial log-link distribution based Model 5 is statistically significant and the model degrees of freedom is seven (7) explanatory variables. The coefficients for miles with no bicycle lane (NBL), miles with speed limit = 45 mph (45mph) and population (POP) are positive. The other explanatory variables such as cul-de-sacs (IT1), dead-ends (IT4), miles with no sidewalk (NSW) and area with single-family (SF) have negative coefficients.

The highest standard error in this Negative Binomial log-link distribution based Model 5 is for the intercept, for which the prediction is off by "0.4347". This is followed

by miles with speed limit = 45 mph (45mph), for which the prediction is off by "0.0468". The least standard error is for single-family (SF) area, for which the prediction is off by "4.2653E-9".

The computed Deviance and Pearson Chi-Square values per degrees of freedom are 1.2 and 1.0, respectively (close to the expected range). The computed AIC = 503.5 and AICC = 505.5 for this model. The estimated dispersion parameter for the model is equal to 0.213.

The developed Negative Binomial log-link distribution based Model 5 is expressed in its exponential form as Equation (10).

$$NC/year = [EXP ((0.159) + (-0.006*IT1) + (-0.015*IT4) + (0.070*NBL) + (-0.033*NSW) + (0.125*45mph) + (-0.00000001075*SF) + (0.00000544*POP))]/6 \quad \text{… Equation (10)}$$

5.7 Negative Binomial Log-link Distribution Based Model 6

Network characteristics that are not correlated to the number of one-way stops on the minor street (IT3), land use characteristics that are not correlated to mixed use (MU) and multi-family (MF) areas, and population are considered initially. Statistically insignificant explanatory variables are eliminated one after another until the model comprises of only significant explanatory variables ($p<=0.05$, at a 95% confidence level). The results obtained for the developed Negative Binomial log-link distribution based Model 6 are summarized and shown in Table 11. The table shows coefficients for each selected explanatory variable (unstandardized coefficient), standard errors (Std. Error), Wald Chi-Square values, p-values (significance values) and goodness-of-fit statistics (Deviance, Pearson Chi-Square, AIC and AICC).

TABLE 11: Negative Binomial log-link distribution based Model 6 - results summary

Parameter	Coeff.	Std. Error	Wald Chi-Square	p-value	D (value/df)	PCS (value/df)	AIC	AICC
(Intercept)	-0.635	0.37	2.90	0.09	1.2	1.0	493.6	496.1
IT1	-0.007	0.00	8.79	<0.01				
IT4	-0.013	0.01	5.71	0.02				
NBL	0.08	0.01	137.59	<0.01				
NSW	-0.035	0.01	14.79	<0.01				
HI	1.42E-08	0.00	5.23	0.02				
MF	4.24E-08	0.00	12.24	<0.01				
RES	6.55E-08	0.00	5.55	0.02				
COM	9.81E-08	0.00	10.02	<0.01				

Note: D = Deviance, PCS = Pearson Chi-Square, AIC = Akaike's Information Criterion
AICC = Finite Sample Corrected AIC

The Omnibus Test indicated that the developed Negative Binomial log-link distribution based Model 6 is statistically significant and the model degrees of freedom is eight (8) explanatory variables. The coefficients for miles with no bicycle lane (NBL), area with heavy industrial (HI), area with multi-family (MF), area with research (RES) and area with commercial (COM) are positive. The other explanatory variables such as cul-de-sacs (IT1), dead-ends (IT4) and miles with no sidewalk (NSW) have negative coefficients.

The highest standard error in this Negative Binomial log-link distribution based Model 6 is for the intercept, for which the prediction is off by "0.3729". This is followed by miles with no sidewalk (NSW), for which the prediction is off by "0.0092". The least standard error is for heavy industrial (HI) area, for which the prediction is off by "6.1997E-9".

The computed Deviance and Pearson Chi-Square values per degrees of freedom are 1.2 and 1.0, respectively (close to the expected range). The computed AIC = 493.640 and AICC = 496.140 for this model. The estimated dispersion parameter for this model is equal to 0.164.

The developed Negative Binomial log-link distribution based Model 6 is expressed in its exponential form as Equation (11).

$$NC/year = [EXP((-0.635) + (-0.007*IT1) + (-0.013*IT4) + (0.080*NBL) + (-0.035*NSW) + (0.00000001418*HI) + (0.00000004236*MF) + (0.00000006549*RES) + (0.00000009813*COM))]/6 \quad \ldots \text{Equation (11)}$$

5.8 Negative Binomial Log-link Distribution Based Model 7

Network characteristics that are not correlated to the number of one-way stops on the minor street (IT3), land use characteristics that are not correlated to single-family (SF) and multi-family (MF) areas, and population are considered initially. Statistically insignificant explanatory variables are eliminated one after another until the model comprises of only significant explanatory variables ($p<=0.05$, at a 95% confidence level). The results obtained for the developed final Negative Binomial log-link distribution based Model 7 are summarized and shown in Table 12. The table shows coefficients for each selected explanatory variable (unstandardized coefficient), standard errors (Std. Error), Wald Chi-Square values, p-values (significance values) and goodness-of-fit statistics (Deviance, Pearson Chi-Square, AIC and AICC).

The Omnibus Test indicated that the developed Negative Binomial log-link distribution based Model 7 is statistically significant and the model degrees of freedom is

five (5) explanatory variables. The coefficients for miles with no bicycle lane (NBL), miles with speed limit = 45 mph (45mph) and area with multi-family (MF) are positive.
The other explanatory variables such as cul-de-sacs (IT1) and miles with no sidewalk (NSW) have negative coefficients.

The highest standard error in this Negative Binomial log-link distribution based Model 7 is for the intercept, for which the prediction is off by "0.3740". This is followed by miles with speed limit = 45 mph (45mph), for which the prediction is off by "0.0470". The least standard error is for area with multi-family (MF) area, for which the prediction is off by "1.2287E-8".

TABLE 12: Negative Binomial log-link distribution based Model 7 - results summary

Parameter	Coeff.	Std. Error	Wald Chi-Square	p-value	D (value/df)	PCS (value/df)	AIC	AICC
(Intercept)	-0.832	0.37	4.94	0.03	1.1	1.0	501.9	503.1
IT1	-0.007	0.00	9.09	<0.01				
NBL	0.078	0.01	109.71	<0.01				
NSW	-0.035	0.01	15.73	<0.01				
45mph	0.107	0.05	5.17	0.02				
MF	4.73E-08	0.00	14.85	<0.01				

Note: D = Deviance, PCS = Pearson Chi-Square, AIC = Akaike's Information Criterion
AICC = Finite Sample Corrected AIC

The computed Deviance and Pearson Chi-Square values per degrees of freedom are 1.1 and 1.0, respectively (within the expected range). The computed AIC = 501.9 and AICC = 503.1 for this model. The estimated dispersion parameter for the model is equal to 0.230.

The developed Negative Binomial log-link distribution based Model 7 is expressed in its exponential form as Equation (12).

NC/year = [EXP ((-0.832) + (-0.007*IT1) + (0.078*NBL) + (-0.035*NSW) + (0.107*45mph) + (0.00000004734*MF))]/6 … Equation (12)

5.9 Negative Binomial Log-link Distribution Based Model 8

Network characteristics that are not correlated to the number of one-way stops on the minor street (IT3), land use characteristics that are not correlated to mixed use (MU) and office (OFF) areas, and population are considered initially. Statistically insignificant explanatory variables are eliminated one after another until the model comprises of only significant explanatory variables (p<=0.05, at a 95% confidence level). The results obtained for the developed final Negative Binomial log-link distribution based Model 8 are summarized and shown in Table 13. The table shows coefficients for each selected explanatory variable (unstandardized coefficient), standard errors (Std. Error), Wald Chi-Square values, p-values (significance values) and goodness-of-fit statistics (Deviance, Pearson Chi-Square, AIC and AICC).

TABLE 13: Negative Binomial log-link distribution based Model 8 - results summary

Parameter	Coeff.	Std. Error	Wald Chi-Square	p-value	D (value/df)	PCS (value/df)	AIC	AICC
(Intercept)	-0.611	0.38	2.58	0.11	1.2	1.0	509.6	510.8
IT1	-0.008	0.00	13.09	<0.01				
NBL	0.07	0.01	55.67	<0.01				
NSW	-0.036	0.01	15.78	<0.01				
45mph	0.155	0.05	10.39	<0.01				
POP	5.82E-05	0.00	6.66	0.01				

Note: D = Deviance, PCS = Pearson Chi-Square, AIC = Akaike's Information Criterion
AICC = Finite Sample Corrected AIC

The Omnibus Test indicated that the developed Negative Binomial log-link distribution based Model 7 is statistically significant and the model degrees of freedom is five (5) explanatory variables. The coefficients for miles with no bicycle lane (NBL), miles with speed limit = 45 mph (45mph) and population (POP) are positive. The other explanatory variables such as cul-de-sacs (IT1) and miles with no sidewalk (NSW) have negative coefficients.

The highest standard error in this Negative Binomial log-link distribution based Model 8 is for the intercept, for which the prediction is off by "0.3803". This is followed by miles with speed limit = 45 mph (45mph), for which the prediction is off by "0.0482". The least standard error is for population (POP), for which the prediction is off by "2.2535E-5".

The computed Deviance and Pearson Chi-Square values per degrees of freedom are 1.2 and 1.0, respectively (close to the expected range). The computed AIC = 509.6 and AICC = 510.8 for this model. The estimated dispersion parameter for this model is equal to 0.257.

The developed Negative Binomial log-link distribution based Model 8 is expressed in its exponential form as Equation (13).

$$NC/year = [EXP ((-0.611) + (-0.008*IT1) + (0.070 *NBL) + (-0.036*NSW) + (0.155*45mph) + (0.00005816*POP)]/6 \quad \ldots \text{Equation (13)}$$

5.10 Model Validation Results

As stated in Chapter 3, each developed model was validated using MFE, MAD, MSE, RMSE, MAPE and SMAPE. Validation dataset for 20 randomly selected locations

that was not used for modeling is considered for validation. Tables 14 to 21 summarize the results obtained from validation of each developed model.

TABLE 14: Model 1 validation results summary

#	Actual # of Crashes (A)	Est. # of Crashes (E)	A-E	Abs(A-E)	$(A-E)^2$	% Error	A+E
1	3	2.0	1.0	1.0	1.0	0.3	5.0
2	1	-0.1	1.1	1.1	1.2	1.1	0.9
3	4	1.0	3.0	3.0	9.3	0.8	5.0
4	2	0.2	1.8	1.8	3.2	0.9	2.2
5	5	2.0	3.0	3.0	9.2	0.6	7.0
6	15	0.7	14.3	14.3	203.4	1.0	15.7
7	6	3.5	2.5	2.5	6.0	0.4	9.5
8	32	3.2	28.8	28.8	829.5	0.9	35.2
9	25	2.9	22.1	22.1	489.8	0.9	27.9
10	7	2.5	4.5	4.5	20.4	0.6	9.5
11	9	0.5	8.5	8.5	72.0	0.9	9.5
12	1	0.1	0.9	0.9	0.8	0.9	1.1
13	3	0.3	2.7	2.7	7.5	0.9	3.3
14	4	1.0	3.0	3.0	8.9	0.7	5.0
15	13	2.5	10.5	10.5	111.2	0.8	15.5
16	2	0.5	1.5	1.5	2.4	0.8	2.5
17	0	-0.9	0.9	0.9	0.7	-	-0.9
18	3	0.9	2.1	2.1	4.3	0.7	3.9
19	5	-0.1	5.1	5.1	25.5	1.0	4.9
20	0	2.3	-2.3	2.3	5.4	-	2.3
	Total		115.0	119.6	1811.6	12.1	165.0
						Number of Samples (N)	20
						Mean Forecast Error (MFE)	5.7
						Mean Absolute Deviation (MAD)	6.0
						Mean Square Error (MSE)	90.6
						Root Mean Square Error (RMSE)	9.5
						Mean Absolute Percentage Error (MAPE)	67.5
						Symmetric Mean Absolute Percent Error (SMAPE)	0.7

TABLE 15: Model 2 validation results summary

#	Actual # of Crashes (A)	Est. # of Crashes	A-E	Abs(A-E)	$(A-E)^2$	% Error	A+E
1	3	3.1	-0.1	0.1	0.0	0.0	6.1
2	1	0.9	0.1	0.1	0.0	0.1	1.9
3	4	2.7	1.3	1.3	1.8	0.3	6.7
4	2	3.6	-1.6	1.6	2.6	0.8	5.6
5	5	9.2	-4.2	4.2	18.1	0.8	14.2
6	15	6.3	8.7	8.7	76.4	0.6	21.3
7	6	15.0	-9.0	9.0	81.6	1.5	21.0
8	32	32.0	0.0	0.0	0.0	0.0	64.0
9	25	8.3	16.7	16.7	279.4	0.7	33.3
10	7	10.8	-3.8	3.8	14.5	0.5	17.8
11	9	6.1	2.9	2.9	8.6	0.3	15.1
12	1	3.5	-2.5	2.5	6.1	2.5	4.5
13	3	3.9	-0.9	0.9	0.8	0.3	6.9
14	4	7.9	-3.9	3.9	15.3	1.0	11.9
15	13	8.1	4.9	4.9	23.7	0.4	21.1
16	2	1.5	0.5	0.5	0.3	0.3	3.5
17	0	0.7	-0.7	0.7	0.5	-	0.7
18	3	2.5	0.5	0.5	0.2	0.2	5.5
19	5	2.8	2.2	2.2	4.7	0.4	7.8
20	0	2.9	-2.9	2.9	8.3	-	2.9
	Total		8.2	67.6	542.8	10.8	271.8
						Number of Samples (N)	20
						Mean Forecast Error (MFE)	0.4
						Mean Absolute Deviation (MAD)	3.4
						Mean Square Error (MSE)	27.1
						Root Mean Square Error (RMSE)	5.2
						Mean Absolute Percentage Error (MAPE)	59.8
						Symmetric Mean Absolute Percent Error (SMAPE)	0.2

TABLE 16: Model 3 validation results summary

#	Actual # of Crashes (A)	Est. # of Crashes (E)	A-E	Abs(A-E)	$(A-E)^2$	% Error	A+E
1	3	2.6	0.4	0.4	0.2	0.1	5.6
2	1	1.4	-0.4	0.4	0.2	0.4	2.4
3	4	2.5	1.5	1.5	2.3	0.4	6.5
4	2	4.7	-2.7	2.7	7.4	1.4	6.7
5	5	5.9	-0.9	0.9	0.9	0.2	10.9
6	15	26.4	-11.4	11.4	129.5	0.8	41.4
7	6	8.0	-2.0	2.0	4.1	0.3	14.0
8	32	25.3	6.7	6.7	44.6	0.2	57.3
9	25	10.9	14.1	14.1	199.8	0.6	35.9
10	7	5.2	1.8	1.8	3.2	0.3	12.2
11	9	14.1	-5.1	5.1	26.0	0.6	23.1
12	1	2.6	-1.6	1.6	2.5	1.6	3.6
13	3	2.6	0.4	0.4	0.1	0.1	5.6
14	4	5.2	-1.2	1.2	1.4	0.3	9.2
15	13	7.6	5.4	5.4	29.6	0.4	20.6
16	2	1.1	0.9	0.9	0.8	0.4	3.1
17	0	0.7	-0.7	0.7	0.6	-	0.7
18	3	1.9	1.1	1.1	1.1	0.4	4.9
19	5	2.5	2.5	2.5	6.4	0.5	7.5
20	0	4.0	-4.0	4.0	16.1	-	4.0
	Total		4.74	64.99	476.89	7.82	275.26
						Number of Samples (N)	20
						Mean Forecast Error (MFE)	0.2
						Mean Absolute Deviation (MAD)	3.2
						Mean Square Error (MSE)	23.8
						Root Mean Square Error (RMSE)	4.9
						Mean Absolute Percentage Error (MAPE)	43.4
						Symmetric Mean Absolute Percent Error (SMAPE)	0.2

TABLE 17: Model 4 validation results summary

#	Actual# of Crashes (A)	Est. # of Crashes (E)	A-E	Abs(A-E)	$(A-E)^2$	% Error	A+E
1	3	2.3	0.7	0.7	0.5	0.2	5.3
2	1	1.2	-0.2	0.2	0.0	0.2	2.2
3	4	2.1	1.9	1.9	3.7	0.5	6.1
4	2	3.8	-1.8	1.8	3.1	0.9	5.8
5	5	7.4	-2.4	2.4	5.9	0.5	12.4
6	15	30.1	-15.1	15.1	227.0	1.0	45.1
7	6	8.1	-2.1	2.1	4.4	0.3	14.1
8	32	24.0	8.0	8.0	63.6	0.2	56.0
9	25	9.7	15.3	15.3	235.3	0.6	34.7
10	7	5.9	1.1	1.1	1.3	0.2	12.9
11	9	7.1	1.9	1.9	3.5	0.2	16.1
12	1	3.5	-2.5	2.5	6.2	2.5	4.5
13	3	4.0	-1.0	1.0	1.0	0.3	7.0
14	4	8.8	-4.8	4.8	23.5	1.2	12.8
15	13	11.8	1.2	1.2	1.4	0.1	24.8
16	2	1.0	1.0	1.0	0.9	0.5	3.0
17	0	0.6	-0.6	0.6	0.3	-	0.6
18	3	2.7	0.3	0.3	0.1	0.1	5.7
19	5	3.2	1.8	1.8	3.1	0.4	8.2
20	0	2.5	-2.5	2.5	6.3	-	2.5
	Total		0.2	66.0	591.0	8.8	279.8

Number of Samples (N)	20
Mean Forecast Error (MFE)	0.0
Mean Absolute Deviation (MAD)	3.3
Mean Square Error (MSE)	29.5
Root Mean Square Error (RMSE)	5.4
Mean Absolute Percentage Error (MAPE)	49.1
Symmetric Mean Absolute Percent Error (SMAPE)	0.2

TABLE 18: Model 5 validation results summary

#	Actual # of Crashes (A)	Est. # of Crashes (E)	A-E	Abs(A-E)	$(A-E)^2$	% Error	SMAPE
1	3	4.5	-1.5	1.5	2.1	0.5	7.5
2	1	1.2	-0.2	0.2	0.1	0.2	2.2
3	4	3.0	1.0	1.0	0.9	0.2	7.0
4	2	4.1	-2.1	2.1	4.2	1.0	6.1
5	5	9.7	-4.7	4.7	21.8	0.9	14.7
6	15	22.2	-7.2	7.2	51.5	0.5	37.2
7	6	9.9	-3.9	3.9	15.2	0.7	15.9
8	32	28.2	3.8	3.8	14.8	0.1	60.2
9	25	11.2	13.8	13.8	191.2	0.6	36.2
10	7	6.9	0.1	0.1	0.0	0.0	13.9
11	9	6.0	3.0	3.0	8.8	0.3	15.0
12	1	3.2	-2.2	2.2	5.1	2.2	4.2
13	3	2.6	0.4	0.4	0.2	0.1	5.6
14	4	7.7	-3.7	3.7	13.6	0.9	11.7
15	13	8.1	4.9	4.9	24.5	0.4	21.1
16	2	1.5	0.5	0.5	0.3	0.3	3.5
17	0	0.7	-0.7	0.7	0.5	-	0.7
18	3	3.0	0.0	0.0	0.0	0.0	6.0
19	5	2.4	2.6	2.6	6.6	0.5	7.4
20	0	3.6	-3.6	3.6	13.2	-	3.6
	Total		0.3	59.8	374.3	8.3	279.7
						Number of Samples (N)	20
						Mean Forecast Error (MFE)	0.0
						Mean Absolute Deviation (MAD)	3.0
						Mean Square Error (MSE)	18.7
						Root Mean Square Error (RMSE)	4.3
						Mean Absolute Percentage Error (MAPE)	45.9
						Symmetric Mean Absolute Percent Error (SMAPE)	0.2

TABLE 19: Model 6 validation results summary

#	Actual# of Crashes (A)	Est. # of Crashes	A-E	Abs(A-E)	$(A-E)^2$	% Error	A+E
1	3	2.3	0.7	0.7	0.4	0.2	5.3
2	1	1.5	-0.5	0.5	0.2	0.5	2.5
3	4	2.3	1.7	1.7	2.8	0.4	6.3
4	2	4.8	-2.8	2.8	7.6	1.4	6.8
5	5	6.1	-1.1	1.1	1.1	0.2	11.1
6	15	30.2	-15.2	15.2	231.3	1.0	45.2
7	6	8.6	-2.6	2.6	6.6	0.4	14.6
8	32	27.2	4.8	4.8	23.1	0.2	59.2
9	25	11.5	13.5	13.5	182.0	0.5	36.5
10	7	5.2	1.8	1.8	3.1	0.3	12.2
11	9	11.4	-2.4	2.4	5.9	0.3	20.4
12	1	2.3	-1.3	1.3	1.7	1.3	3.3
13	3	2.8	0.2	0.2	0.0	0.1	5.8
14	4	5.2	-1.2	1.2	1.5	0.3	9.2
15	13	7.7	5.3	5.3	27.6	0.4	20.7
16	2	1.0	1.0	1.0	1.0	0.5	3.0
17	0	0.7	-0.7	0.7	0.4	-	0.7
18	3	2.1	0.9	0.9	0.9	0.3	5.1
19	5	2.4	2.6	2.6	6.6	0.5	7.4
20	0	3.5	-3.5	3.5	12.5	-	3.5
	Total		1.1	63.6	516.7	7.6	278.9
						Number of Samples (N)	20
						Mean Forecast Error (MFE)	0.1
						Mean Absolute Deviation (MAD)	3.2
						Mean Square Error (MSE)	25.8
						Root Mean Square Error (RMSE)	5.1
						Mean Absolute Percentage Error (MAPE)	42.0
						Symmetric Mean Absolute Percent Error (SMAPE)	0.2

TABLE 20: Model 7 validation results summary

#	Actual# of Crashes (A)	Est. # of Crashes	A-E	Abs(A-E)	$(A-E)^2$	% Error	A+E
1	3	3.0	0.0	0.0	0.0	0.0	6.0
2	1	1.4	-0.4	0.4	0.2	0.4	2.4
3	4	2.0	2.0	2.0	4.1	0.5	6.0
4	2	3.0	-1.0	1.0	1.0	0.5	5.0
5	5	6.5	-1.5	1.5	2.4	0.3	11.5
6	15	36.7	-21.7	21.7	471.1	1.4	51.7
7	6	14.7	-8.7	8.7	75.9	1.5	20.7
8	32	24.8	7.2	7.2	52.1	0.2	56.8
9	25	10.0	15.0	15.0	224.3	0.6	35.0
10	7	6.4	0.6	0.6	0.4	0.1	13.4
11	9	6.5	2.5	2.5	6.2	0.3	15.5
12	1	2.8	-1.8	1.8	3.1	1.8	3.8
13	3	2.3	0.7	0.7	0.5	0.2	5.3
14	4	5.3	-1.3	1.3	1.7	0.3	9.3
15	13	7.7	5.3	5.3	28.2	0.4	20.7
16	2	1.3	0.7	0.7	0.5	0.4	3.3
17	0	0.7	-0.7	0.7	0.5	-	0.7
18	3	2.9	0.1	0.1	0.0	0.0	5.9
19	5	2.2	2.8	2.8	7.8	0.6	7.2
20	0	1.6	-1.6	1.6	2.5	-	1.6
	Total		-1.7	75.8	882.6	8.6	281.7
						Number of Samples (N)	20
						Mean Forecast Error (MFE)	-0.1
						Mean Absolute Deviation (MAD)	3.8
						Mean Square Error (MSE)	44.1
						Root Mean Square Error (RMSE)	6.6
						Mean Absolute Percentage Error (MAPE)	47.9
						Symmetric Mean Absolute Percent Error (SMAPE)	0.3

TABLE 21: Model 8 validation results summary

#	Actual# of Crashes (A)	Est. # of Crashes	A-E	Abs(A-E)	$(A-E)^2$	% Error	A+E
1	3	4.5	-1.5	1.5	2.2	0.5	7.5
2	1	1.8	-0.8	0.8	0.6	0.8	2.8
3	4	2.9	1.1	1.1	1.3	0.3	6.9
4	2	3.3	-1.3	1.3	1.6	0.6	5.3
5	5	8.1	-3.1	3.1	9.5	0.6	13.1
6	15	29.5	-14.5	14.5	210.7	1.0	44.5
7	6	12.9	-6.9	6.9	48.1	1.2	18.9
8	32	29.3	2.7	2.7	7.2	0.1	61.3
9	25	10.8	14.2	14.2	200.5	0.6	35.8
10	7	7.9	-0.9	0.9	0.8	0.1	14.9
11	9	6.4	2.6	2.6	6.9	0.3	15.4
12	1	3.2	-2.2	2.2	4.9	2.2	4.2
13	3	2.4	0.6	0.6	0.3	0.2	5.4
14	4	6.4	-2.4	2.4	5.6	0.6	10.4
15	13	5.9	7.1	7.1	49.9	0.5	18.9
16	2	1.5	0.5	0.5	0.2	0.2	3.5
17	0	1.0	-1.0	1.0	1.0	-	1.0
18	3	3.8	-0.8	0.8	0.6	0.3	6.8
19	5	2.6	2.4	2.4	5.9	0.5	7.6
20	0	2.0	-2.0	2.0	4.0	-	2.0
	Total		-6.2	68.5	561.9	9.3	286.2

Number of Samples (N)	20
Mean Forecast Error (MFE)	-0.3
Mean Absolute Deviation (MAD)	3.4
Mean Square Error (MSE)	28.1
Root Mean Square Error (RMSE)	5.3
Mean Absolute Percentage Error (MAPE)	51.7
Symmetric Mean Absolute Percent Error (SMAPE)	0.2

5.10.1 Model 1 Validation Interpretations

The results obtained from validation of Model 1 are summarized in this subsection.

(1) The computed MFE for Model 1 is 5.7. This indicates the model's tendency to underestimate i.e., the actual number of bicycle-vehicle crashes is more than the estimated number of bicycle-vehicle crashes.

(2) The computed MAD and MSE are 6.0 and 90.6, respectively. They cannot be considered on their own and need to be compared to MAD and MSE of other models or standards. These two validation test parameters are called "comparative numbers". In comparison of models, the lower the MSE or MAD, the better or close to actual estimates.

(3) The computed RMSE is 9.5, while the computed MAPE is 67.5. MAPE is a measure of prediction accuracy of a forecasting method in statistics, which expresses accuracy as a percentage. MAPE has issues arising from its practical applications. It cannot be used if there are zero values in the actual number of bicycle-vehicle crashes (because there would be a division by zero). When MAPE is used to compare the accuracy of prediction methods, it is biased in that it systematically selects a method whose forecasts are too low. In addition, it puts a heavier penalty on negative errors ($A < E$) than on positive errors. To overcome these issues with MAPE, SMAPE or Mean Absolute Scaled Error (MASE) are used.

(4) The computed SMAPE is 0.7. It is an accuracy measure based on percentage errors. It measures the direction of bias in the data by generating a positive and a negative error on line item level. SMAPE is better protected against outliers and the bias effect. The limitation to SMAPE is that if the actual value or estimated (forecast) value is zero (0),

the value of error will boom up to the upper limit of error. Having two zeros in the actual number of bicycle-vehicle crashes does not guaranty a good result from SMAPE neither.

5.10.2 Model 2 Validation Interpretations

The results obtained from validation of Model 2 are summarized in this subsection.

(1) The computed MFE is 0.4 for Model 2. Even this model has the tendency to underestimate.

(2) The computed MAD and MSE are 3.4 and 27.1, respectively.

(3) The computed RMSE, MAPE and SMAPE are 5.2, 59.8 and 0.2, respectively.

5.10.3 Model 3 Validation Interpretations

The results obtained from validation of Model 3 are summarized in this subsection.

(1) The computed MFE is 0.2 for Model 3. Even this model has the tendency to underestimate.

(2) The computed MAD and MSE are 3.2 and 23.8, respectively.

(3) The computed RMSE, MAPE and SMAPE are 4.9, 43.4 and 0.2, respectively.

5.10.4 Model 4 Validation Interpretations

The results obtained from validation of Model 4 are summarized in this subsection.

(1) The computed MFE is 0.0 for Model 4. It is lowest of all the developed models.

(2) The computed MAD and MSE are 3.3 and 29.5, respectively.

(3) The computed RMSE, MAPE and SMAPE are 5.4, 49.1 and 0.2, respectively.

5.10.5 Model 5 Validation Interpretations

The results obtained from validation of Model 5 are summarized in this subsection.

(1) The computed MFE is 0.0 for Model 5 as well.

(2) The computed MAD and MSE are 3.0 and 18.7, respectively.

(3) The computed RMSE, MAPE and SMAPE are 4.3, 45.9 and 0.2, respectively.

5.10.6 Model 6 Validation Interpretations

The results obtained from validation of Model 7 are summarized in this subsection.

(1) The computed MFE is 0.1 for Model 6. Even this model has the tendency to underestimate.

(2) The computed MAD and MSE are 3.2 and 25.8, respectively.

(3) The computed RMSE, MAPE and SMAPE are 5.1, 42.0 and 0.2, respectively.

5.10.7 Model 7 Validation Interpretations

The results obtained from validation of Model 7 are summarized in this subsection.

(1) The computed MFE is -0.1 for Model 7. This model has the tendency to overestimate.

(2) The computed MAD and MSE are 3.8 and 44.1, respectively.

(3) The computed RMSE, MAPE and SMAPE are 6.6, 47.9 and 0.3, respectively.

5.10.8 Model 8 Validation Interpretations

The results obtained from validation of Model 8 are summarized in this subsection.

(1) The computed MFE is -0.3 for Model 8. This model has the tendency to overestimate as well.

(2) The computed MAD and MSE are 3.4 and 28.1, respectively.

(3) The computed RMSE, MAPE and SMAPE are 5.3, 51.7 and 0.2, respectively.

5.11 Comparison and Selection of the Best Model

Table 22 summarizes the goodness-of-fit statistics for all the developed eight models, while Table 23 summarizes model validation results for all the developed eight models.

The computed Deviance and Pearson Chi-Square values per degrees of freedom are highest for Model 1, followed by Model 2. It is lowest for Model 7 (an indicator of low residuals). The Pearson Chi-Square values per degrees of freedom is equal to 1.0 for models 3 to 8, when explanatory variables that are not correlated to each other were considered for modeling. The AIC and AICC are lowest for Model 2, followed by Model 3 and Model 6.

TABLE 22: Summary of model goodness-of-fit statistics

Model #	Deviance (value/df)	Pearson Chi-Square (value/df)	AIC	AICC
1	1.9	1.8	498.4	659.7
2	1.3	1.2	475.9	480.9
3	1.2	1.1	491.8	495.4
4	1.2	1.0	501.4	503.0
5	1.2	1.0	503.5	505.5
6	1.2	1.0	493.6	496.1
7	1.1	1.0	501.9	503.1
8	1.2	1.0	509.6	510.8

TABLE 23: Summary of model validation results

Model #	MFE	MAD	MSE	RMSE	MAPE	SMAPE
1	5.7	6.0	90.6	9.5	67.5%	70.0%
2	0.4	3.4	27.1	5.2	59.8%	20.0%
3	0.2	3.2	23.8	4.9	43.4%	20.0%
4	0.0	3.3	29.5	5.4	49.1%	20.0%
5	0.0	3.0	18.7	4.3	45.9%	20.0%
6	0.1	3.2	25.8	5.1	42.0%	20.0%
7	-0.1	3.8	44.1	6.6	47.9%	30.0%
8	-0.3	3.4	28.1	5.3	51.7%	20.0%

The computed MFE, MAD, MSE, RMSE, MAPE and SMAPE are lowest for Model 5, followed by Model 3. The models 1, 2, 3, and 6 seem to underestimate bicycle-vehicle crash frequency, while models 7 and 8 seem to overestimate bicycle-vehicle crash frequency. The MFE is 0 for models 4 and 5.

The computed MAD values (mean of errors without the basis of overestimating or underestimating) indicates that Model 5 is off by an average of 3.0 units, Model 3 and 6 are off by an average of 3.2 units, Model 4 is off by an average of 3.3 units, Model 2 and 8 are off by an average of 3.4 units, Model 7 is off by an average of 3.8 units and Model 1 off by an average of 6.0 units (highest) from the actual bicycle-vehicle crash frequency.

On the other hand, MSE penalizes large errors by giving weights when validating. Model 5 has the lowest MSE of 18.7, followed by Model 3 with 23.8. Likewise, RMSE is lowest for Model 5, followed by Model 3. Model 1 has the highest MSE and RMSE.

The computed RMSE values range from 4.3 to 9.5. These values seem to be high, greater than the mean bicycle-vehicle crash frequency in some cases. Overestimating or underestimating bicycle-vehicle crash frequency at locations with higher bicycle-vehicle crash frequency (> 20 bicycle-vehicle crashes) and congested traffic conditions is observed to be the primary problem. This can be observed from the spatial pattern of computed residuals shown in figures 17 to 24. To further assist with the interpretation of spatial patterns in residuals, Global Moran's Index was computed in GIS environment to examine if the spatial pattern of residuals from each model is clustered, dispersed, or random. An inverse distance concept was felt appropriate and used for analysis. When the z-score indicates statistical significance, a Moran's Index value close to +1.0 indicates clustering

while a Moran's Index value close to –1.0 indicates dispersion. The results obtained are summarized in Table 24.

TABLE 24: Summary of Moran's Index, z-score and p-values

Model #	Moran's Index	z-score	p-value
Model 1	0.05	5.03	<0.01
Model 2	-0.01	-0.27	0.79
Model 3	-0.03	-1.53	0.13
Model 4	-0.03	-1.73	0.08
Model 5	-0.03	-1.59	0.11
Model 6	-0.03	-1.36	0.18
Model 7	-0.02	-1.01	0.31
Model 8	-0.04	-2.06	0.04

The z-score of Model 1 indicates that the spatial pattern of its residuals is clustered at a 99% confidence level. On the other hand, the z-scores of Model 8 and Model 4 indicate that the spatial pattern of their residuals is dispersed at 95% and 90% confidence level, respectively. For all other models, the spatial pattern does not appear to be significantly different than random (possibly attributed to the selection of locations for modeling). In general, the computed Moran's Index values are close to 0, indicating weak to no spatial correlation between residuals of each model.

Comparing parameters and validation results for models 3 to 8 with models 1 and 2 indicates that considering explanatory variables that are not correlated to each other and eliminating statistically insignificant variables improve the predictability of bicycle-vehicle crash frequency models. Further, the number of explanatory variables (hence, data collection efforts) could be minimal if models 3 to 8 are used.

Based on results from validation, Model 5 could be classified as a good model. However, it does not consider critical variables such as traffic lights (directly related to conflicting situations), high speed urban roads, research, commercial and land use mix such as multi-family and uptown mixed use areas that were identified as critical explanatory variables by past researchers.

Model 3, on the other hand, seem to perform more consistently based on statistical parameters and validation results and is, therefore, recommended as the model to estimate bicycle-vehicle crash frequency. Additional justification for selection of Model 3 is presented next.

(1) First, many previous studies observed that intersections are critical locations that play a major role in the number of reported traffic crashes (Wachtel and Lewiston, 1994); Klop and Khattak, 1999; Delmelle and Thill, 2008; Reynolds et al., 2009; Pulugurtha and Repaka, 2011; Wei and Lovegrove, 2012; Strauss et al., 2013; Pulugurtha and Imran, 2013; Wang et al., 2013; Figliozzi et al., 2013). In particular, traffic lights are major contributors of bicycle-vehicle crashes (Pulugurtha & Thakur, 2015). This explanatory variable (IT5) is included in Model 3.

(2) Second, bicyclists are three to four times at higher risk (based on traffic conditions) while traveling on segments without on-street bicycle lane than on segments with on-street bicycle lane (Pulugurtha & Thakur, 2015). Reynolds et al (2009) indicated that on-road marked bicycle lane is associated with the lowest risk for bicyclists. This factor is captured as center-line miles without bicycle lane (NBL) and included in Model 3.

(3) Third, past research affirmed that institutional areas (e.g. schools), research, business districts, commercial, employment, local city streets, sidewalks, and land use mix

increase risk to bicyclists (Wachtel and Lewiston, 1994; Kim et al., 2007; Delmelle and Thill, 2008; Reynolds et al., 2009; Delmelle et al., 2012; Strauss et al., 2013). Several of these explanatory variables are included in Model 3.

The absence of sidewalk decreases bicycle-vehicle crash frequency in such way that the bicyclist is less distracted by the presence of pedestrians on the sidewalk. Bicyclists often make sudden intrusion on the vehicle path in an effort to avoid the collision with a pedestrian on the sidewalk. The unevenness of travel path (example, at intersection with ADA ramps) only aggravates the problem. Dead-ends (IT4) and cul-de-sacs (IT1) have relatively lower number of conflicts, while the risk of getting involved in a bicycle-vehicle crash is high at traffic lights with relatively more number of conflicts. The absence of bicycle lane exposes bicyclists to additional risk. Traffic volumes and speeds are higher in heavy industrial, research and commercial areas (increasing risk of bicyclist getting involved in a crash), while bicycling activity is typically higher in multi-family and mixed use areas. Therefore, using Model 3 will help better forecast bicycle-vehicle crashes on urban roads. This forecasting model helps identify appropriate solutions and proactively improve safety of bicycle riders in urban settings.

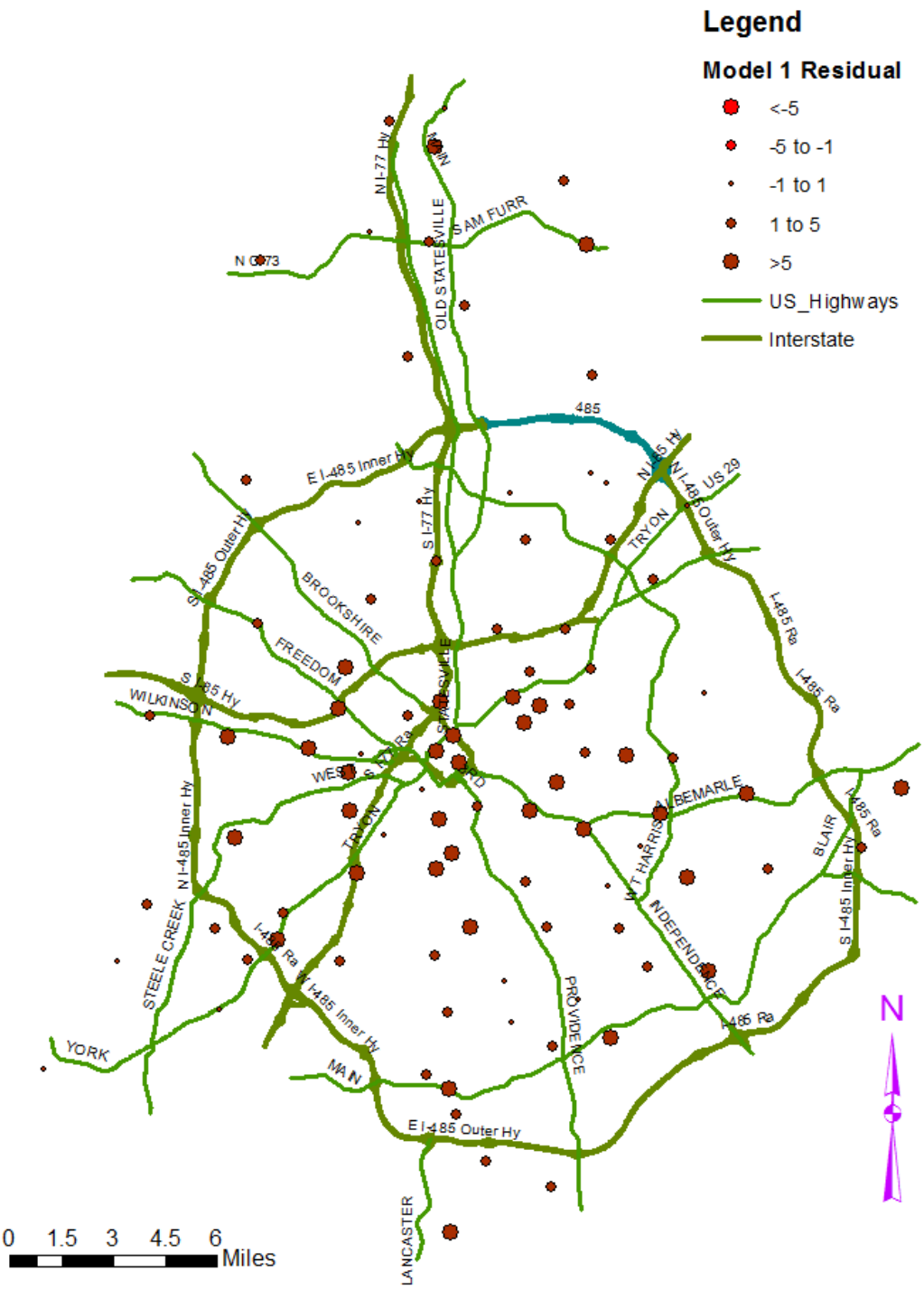

FIGURE 17: Model 1 residuals - spatial pattern

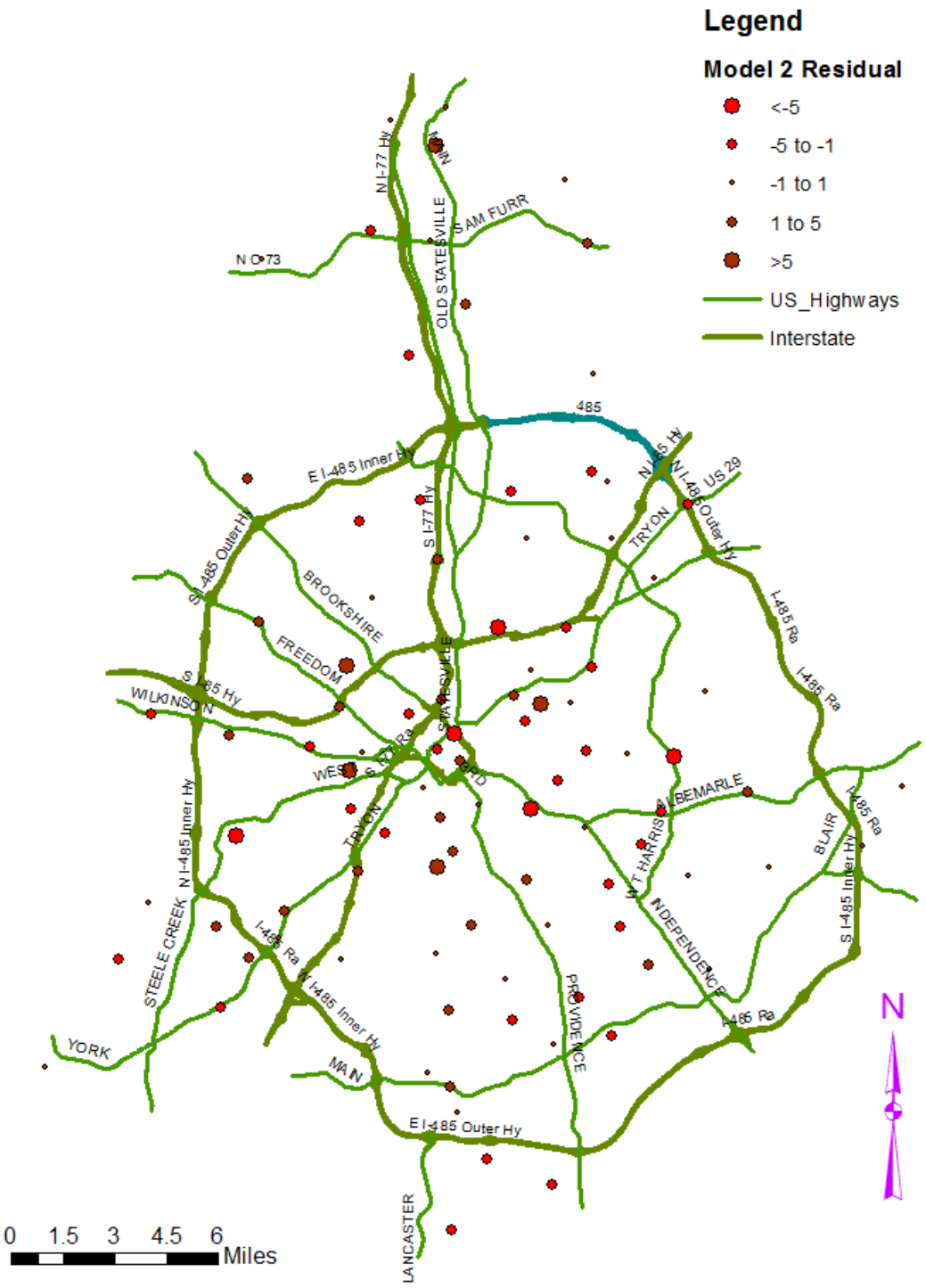

FIGURE 18: Model 2 residuals - spatial pattern

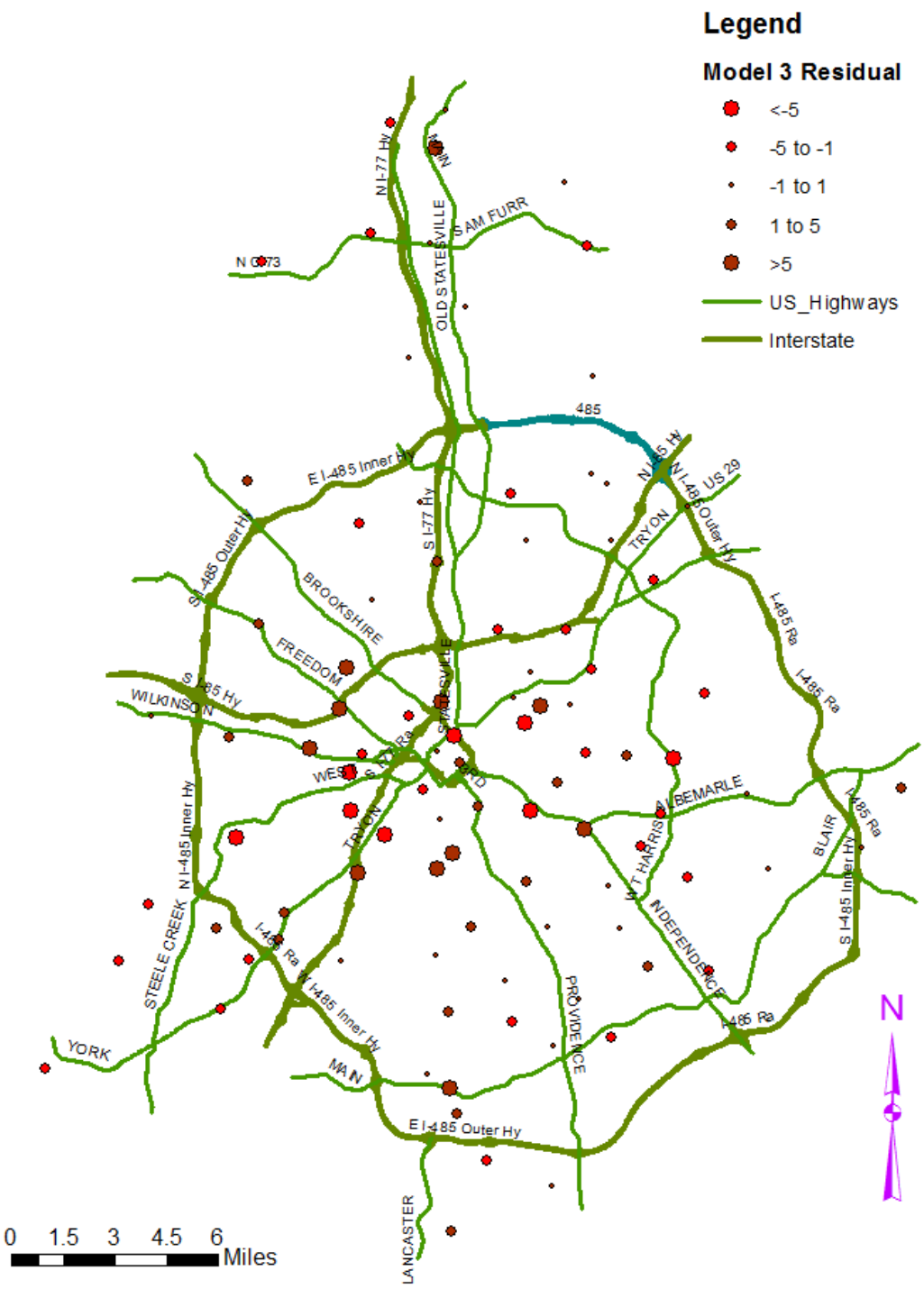

FIGURE 19: Model 3 residuals - spatial pattern

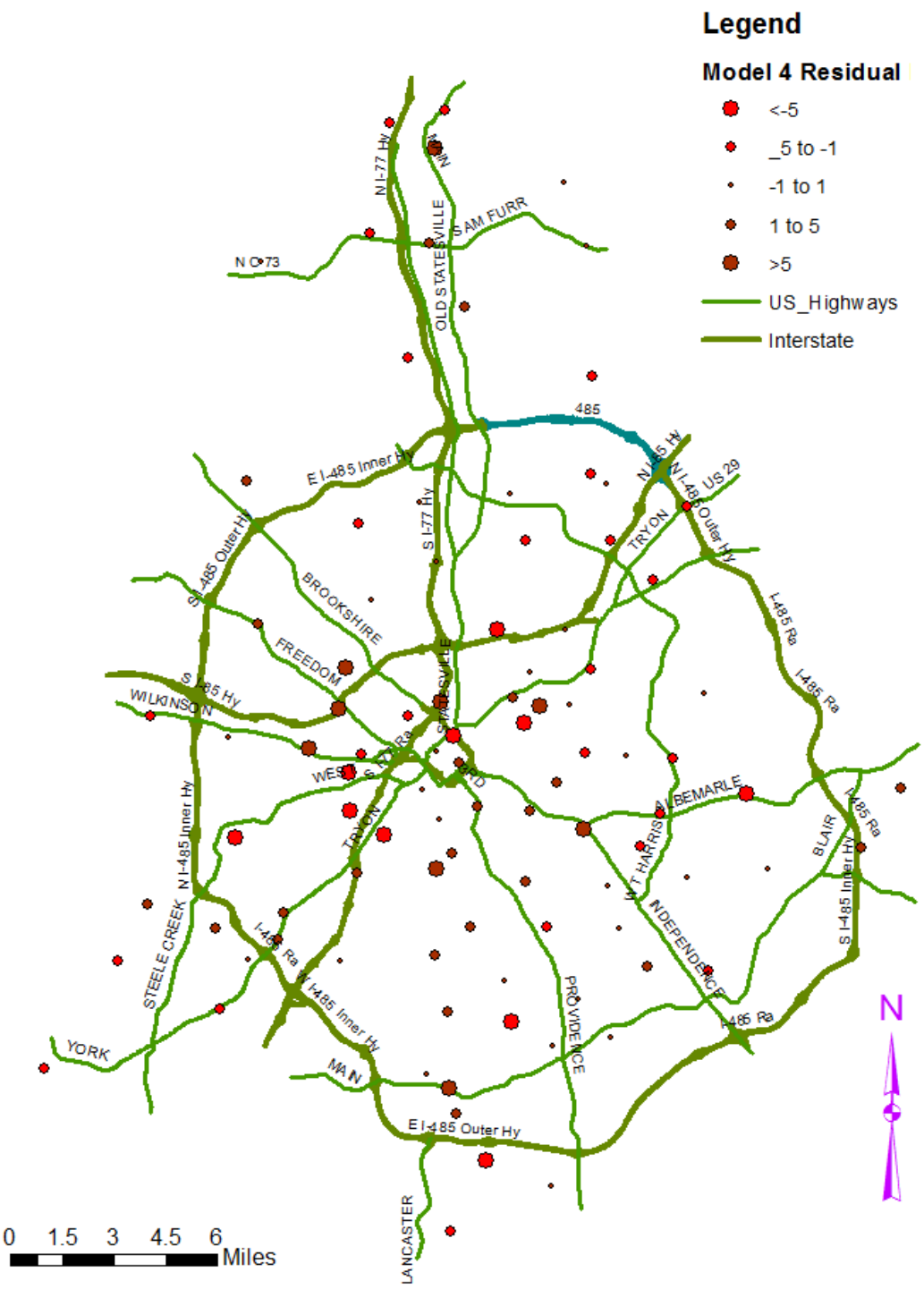

FIGURE 20: Model 4 residuals - spatial pattern

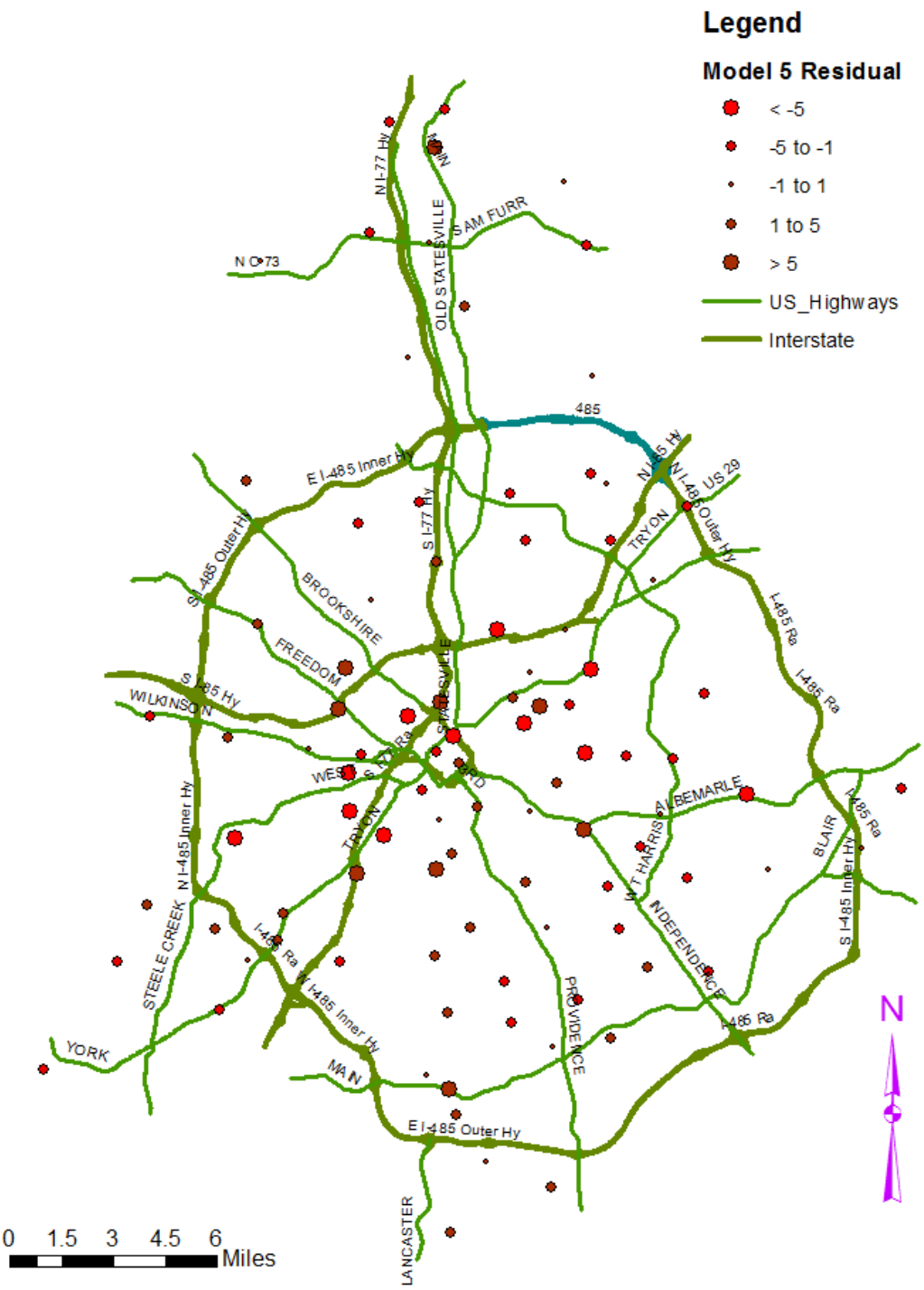

FIGURE 21: Model 5 residuals - spatial pattern

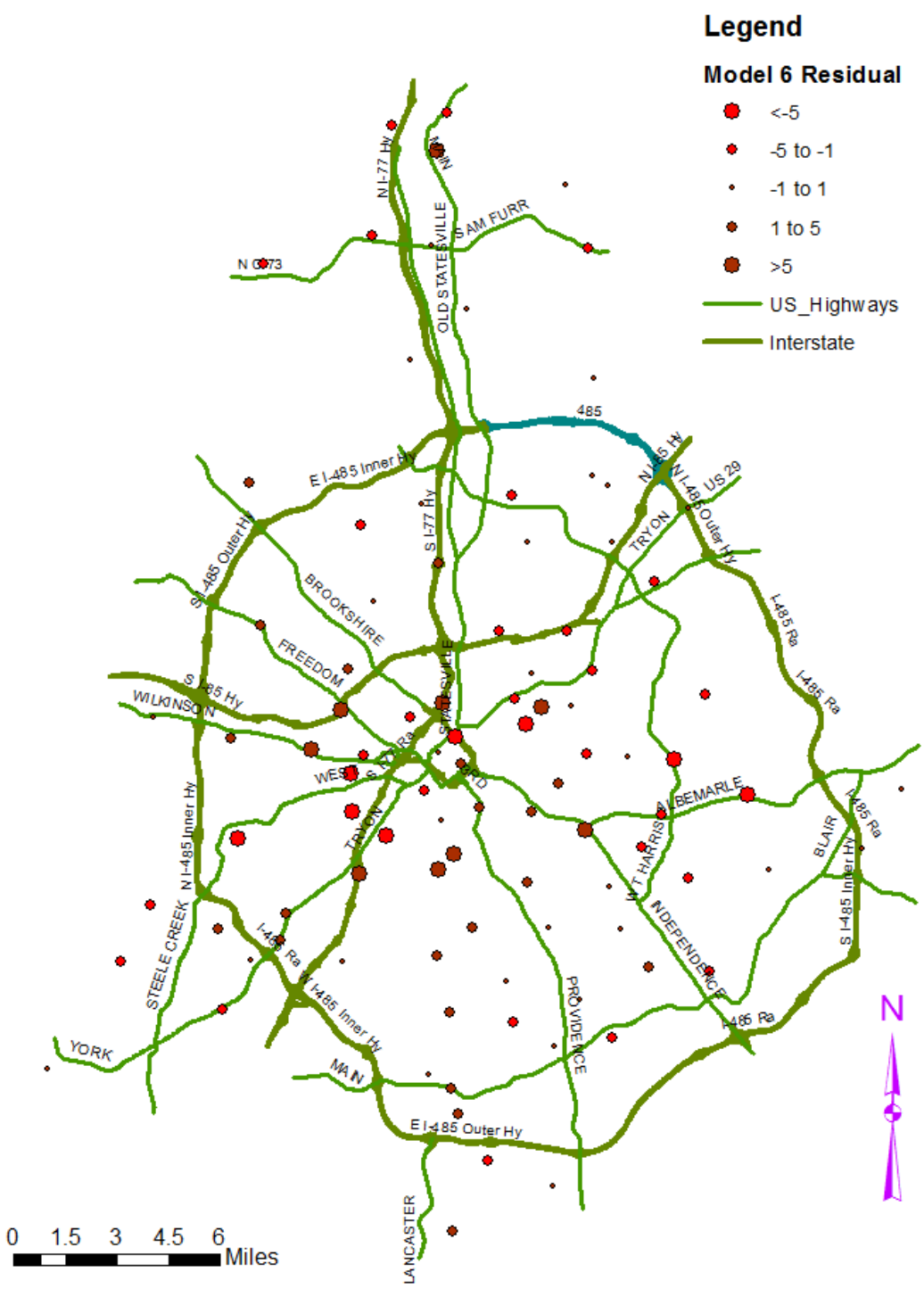

FIGURE 22: Model 6 residuals - spatial pattern

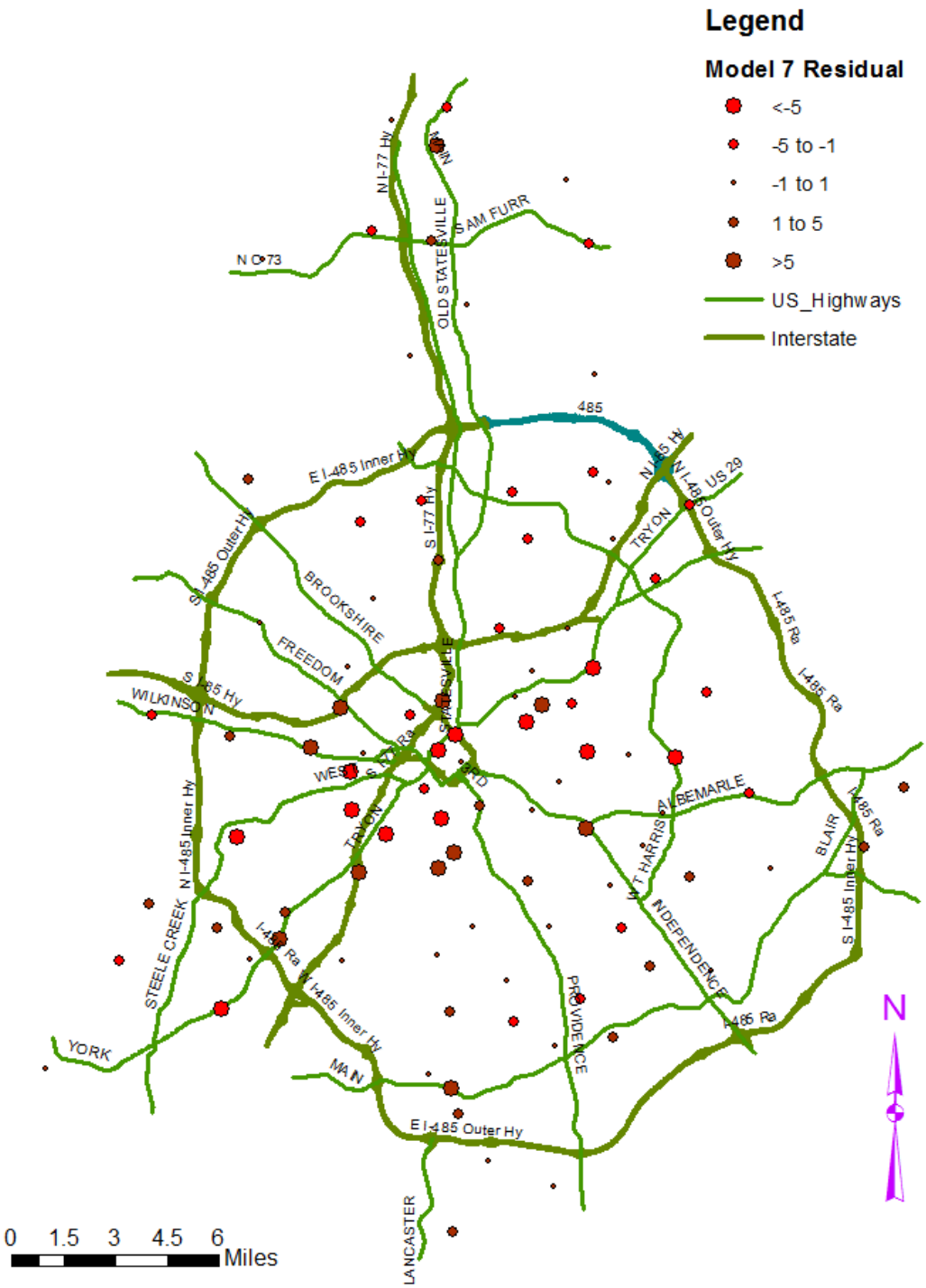

FIGURE 23: Model 7 residuals - spatial pattern

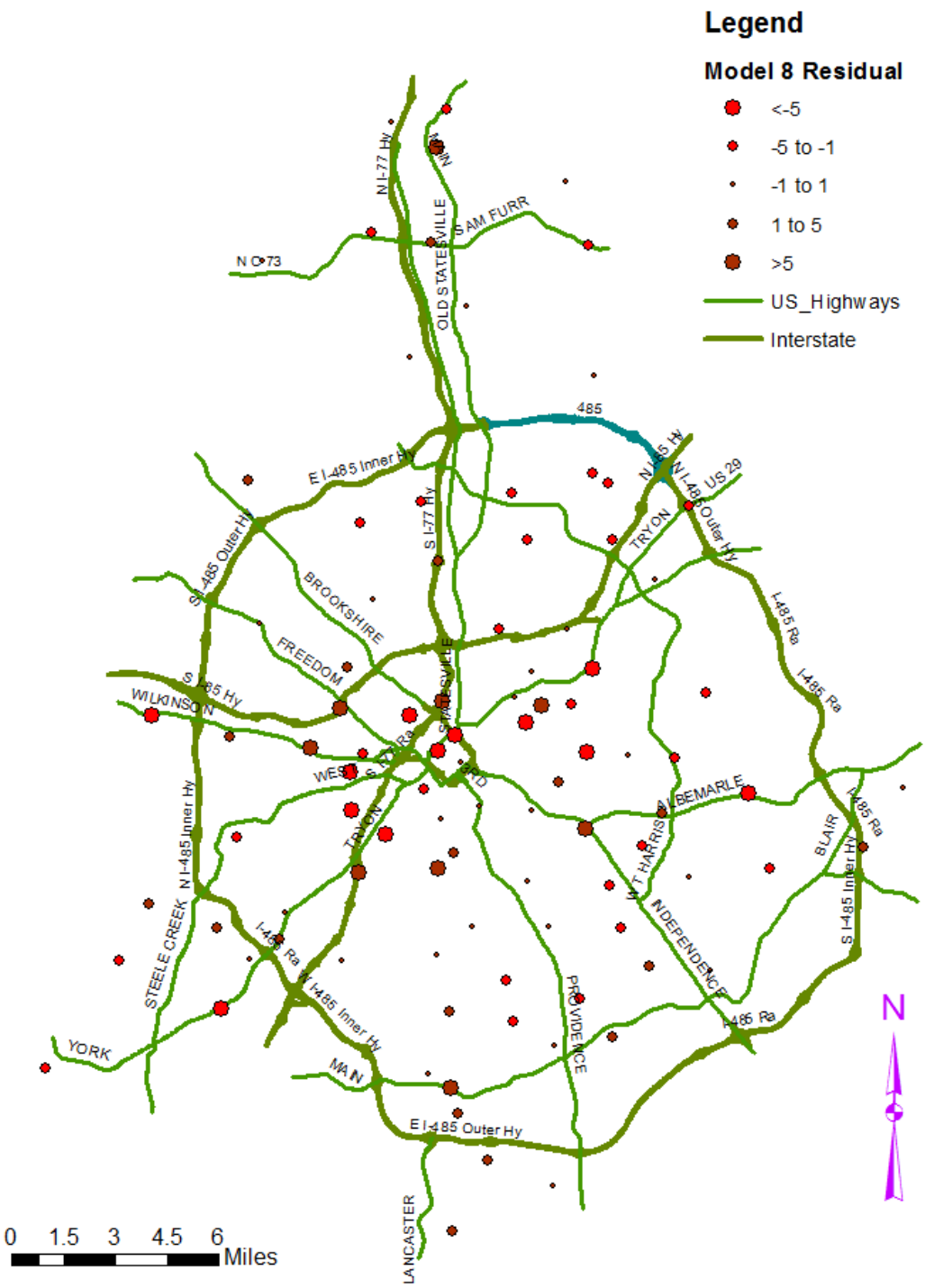

FIGURE 24: Model 8 residuals - spatial pattern

CHAPTER 6: CONCLUSIONS

In this research, macroscopic bicycle-vehicle crash frequency models were developed with emphasis on demographic, land use, and network characteristics. The methodology adopted used tools available in GIS, data analytics and statistical methods to identify explanatory variables and estimate bicycle-vehicle crash frequency (safety performance function).

Selected demographic characteristics, land use characteristics and various network characteristics (overall, 55 explanatory variables) were captured in GIS environment for 119 locations. These locations account for 91.8% of the observed bicycle-vehicle crashes during the study period. The selected locations are geographically distributed and fall in high, medium, low and no risk locations.

Data for 99 randomly selected locations was used to develop bicycle-vehicle crash frequency models, while data for 20 randomly selected locations was used to validate the models. These macroscopic models estimate the overall bicycle-vehicle crash frequency in an area within a radius equal to 1-mile. The results from analysis and models can assist planners, professionals, practitioners and policy-makers to correct the hazards, to develop rezoning plans, and understand the role of developments for improving the safety of bicyclists on urban roads.

Results obtained from “One Sample Kolmogorov-Smirnov Test” and “Descriptive Statistics” indicate that the bicycle-vehicle crash data used in this research is over-dispersed. Therefore, Negative Binomial log-link distribution based models were

developed to estimate bicycle-vehicle crash frequency than Poisson log-link distribution based count models.

Strong correlations were observed between bicycle-vehicle crash frequency on urban roads and most network, land use, and demographic characteristics within a 1-mile radius. With exceptions of the number of cul-de-sacs (IT1), mixed use residential (MUR) area and single-family (SF) residential area, an increase in all other considered significant explanatory variables could lead to an increase in bicycle-vehicle crash frequency.

The computed Pearson correlation coefficients indicate that strong correlations exist between selected network, land use and demographic characteristics (p-value =~0.000). Bicycle-vehicle crash frequency models were, therefore, developed with and without considering explanatory variables that are strongly correlated to each other, and, with and without eliminating statistically insignificant explanatory variables.

The results obtained from analysis and modeling indicate that bicyclists are at a significantly higher risk of getting involved in a crash while traveling

(1) on segments with no bicycle lane,

(2) on segments with traffic lights,

(3) on segments with 45 mph as speed limit,

(4) in commercial areas,

(5) in areas with research activity and institutions,

(6) in areas with multi-family residential units (densely populated), and,

(7) in heavy industrial areas.

These could be associated with exposure (conflicting situations, high speed / high volume roads, etc.) and areas with high bicycling activity levels. On the other hand, cul-de-sacs, dead-ends and single-family residential area could have a lower but smoothing effect on bicycle-vehicle crash frequency. Overall, based on the results and observed magnitude of forecasting errors, network characteristics have equal or better predictive ability than land use and demographic characteristics considered in this research.

Not considering two explanatory variables that are correlated to each other and eliminating insignificant explanatory variables when developing bicycle-vehicle crash frequency models tend to improve the predictability of the models. This is well supported by results obtained from model validation.

The developed methodology and bicycle-vehicle crash frequency models (safety performance functions) can be used to estimate bicycle-vehicle crash frequency within the vicinity of any location and proactively incorporate bicyclist's safety into land use decisions, transportation improvement programs, metropolitan transportation plans and comprehensive transportation plans to minimize projected bicycle-vehicle crash frequency.

6.1 Limitations and Scope for Further Research

Only 22 fatal and severe injury (Type A) bicycle-vehicle crashes were observed in the study area during the study period. This sample is inadequate to develop bicycle-vehicle crash frequency models by severity. Data from multiple study areas should be combined to develop and test the validity of bicycle-vehicle crash frequency models by severity.

The residuals seem to be relatively high for locations with higher number of bicycle-vehicle crash frequency. The bicycle-vehicle crash frequency is typically higher in and around uptown/downtown area. Developing bicycle-vehicle crash frequency models by area type (central business district, urban and suburban areas) might lead to accurate estimates. Large data from multiple study areas should be gathered and used to develop and validate such bicycle-vehicle crash frequency models by area type.

Explanatory variables such as bicycle counts, AADT, gender, and age-group were not considered for analysis and modeling. Bicycle counts are not available for the study area, while AADT was not available for collector and local roads that constitute the majority of urban roads in this study. Likewise, gender and age was not available in the obtained bicycle-vehicle crash data. Additionally, implementation of bicycle racks and rentals is growing. Collecting and considering such data for analysis and modeling not only improves predictability and understand the role of causal factors but also helps to identify solutions and strategies that enhance safety of bicyclists on urban roads. This would certainly benefit from incorporating bicycle counts and expanding traffic counts as a part traffic data collection programs.

The accuracy of estimates from the models depend on the accuracy of data used for developing the models. Likewise, collecting the data regularly and adopting consistent standards to maintain the data is key when data from multiple study areas are used for modeling.

REFERENCES

Abdel-Aty, M. A., C. L. Chen, and J. R. Schott. 1998. An assessment of the effect of driver age on traffic accident involvement using log-linear. Accident Analysis & Prevention Journal 30(6): 851-861.

Abdel-Aty, M. A. and A.E. Radwan. 2000. Modeling traffic accident occurrence and involvement. Accident Analysis & Prevention Journal 32: 633-642.

Amoh-Gyimah, R., M. Sarvi, and M. Saberi. 2016. Investigating the effects of traffic, socioeconomic, and land use characteristics on pedestrian and bicycle crashes: a case study of Melbourne, Australia. Transportation Research Board 95th Annual Meeting (TRB Paper # 16-1931), Washington, DC.

Anastasopoulos, P. Ch., and F. L. Mannering. 2008. A note on modeling vehicle accident frequencies with random-parameters count model. Accident Analysis & Prevention Journal 41:153-159.

Anderson, T. 2009. Kernel density estimation and K-means clustering to profile road accident hotspots. Accident Analysis & Prevention Journal 41: 359-364.

Armstrong, J. and Z. Petch. 2013. School transport walking hazard assessment guidelines. Transportation Association of Canada, Winnipeg, Manitoba.

Beimborn, E. 1999. An overview: land use and economic development in statewide transportation planning. TRB Statewide Transportation Planning Conference 7/21/99.

Bin Islam, M. and S. Hernandez. 2013. Modeling injury outcomes of crashes involving heavy vehicles on Texas highways. Transportation Research Record Journal 2388: 28-36.

Bolstad, P. (2012). GIS fundamentals: a first text on geographic information systems. 4th Edition, Eider Press: Minnesota, MN.

Chaurand, N. and P. Delhomme. 2013. Cyclists and drivers in road interactions: a comparison of perceived crash risk. Journal of Accident Analysis & Prevention Journal 50: 1176-1184.

Chimba, D., D. Emaasit, C. R. Cherry, and Z. Pannell. 2014. Patterning demographic and socioeconomic characteristics affecting pedestrian and bicycle crash frequency.

Transportation Research Board 93rd Annual Meeting (TRB Paper # 14-0600), Washington, DC.

Chiou, Y. and C. Fu. 2013. Modeling crash frequency and severity using multinomial-generalized Poisson model with error components. Accident Analysis & Prevention Journal 50: 73-82.

Delmelle, E.C. and J.-C. Thill. 2008. Urban bicyclists: spatial analysis of adult and youth traffic hazard intensity. Transportation Research Record Journal 2074: 31-39.

Delmelle, E. C., J.-C. Thill, and E. Delmelle. (2008). Using GIS to determine risk factors of urban bicyclists: Buffalo, NY case study. ResearchGate.net/publication/241681609.

Delmelle, E.C., J.-C. Thill, and H.-H. Ha. 2012. Spatial epidemiologic analysis of relative collision risk factors among urban bicyclists and pedestrians. Transportation Research Board 91st Annual Meeting, Washington, DC.

Delmelle, E. C. 2016. Mapping the DNA of urban neighborhoods: clustering longitudinal sequences of neighborhood socioeconomic change. Annals of the American Association of Geographers, 106(1): 36-56.

Emaasit, D., D. Chimba, C.R. Cherry, B. Kutela, and J. Wilson. 2013. A methodology to identify factors associated with pedestrian high crash clusters using GIS based local spatial autocorrelation. Transportation Research Board 92nd Annual Meeting (TRB Paper #13-0634), Washington, DC.

Falb, M., D. Kanny, K. Powell, and A. Giarrusso. 2007. Estimating the proportion of children who can walk to school. American Journal of Preventive Medicine 33(4): 269-275.

Fan, W., M. Kane, and E. Haile. 2015a. Predicting the severity of pedestrian crashes on highway-rail grade crossings. Advances in Transportation Studies, an International Journal Section A 36: 63-74.

Fan, W., M. Kane, and E. Haile. (2015b). Analyzing severity of vehicle crashes at highway-rail grade crossings: multinomial logit modeling. Journal of the Transportation Research Forum 54(2):39-54.

Farley, H. and Z. Smith. 2014. Sustainability: if it's everything is it nothing? Routeledge Taylor & Francis Group, First Edition, New York, NY.

Figliozzi, M., N. Wheeler, and C. Monsere. 2013. A methodology to estimate bicyclists' acceleration and speed distributions at signalized intersections. Transportation Research Board 92nd Annual Meeting (TRB Paper #13-2697), Washington, DC.

Fils, P. B. 2012. Modeling travel time and reliability on urban arterials for recurrent conditions. Graduate Theses and Dissertations. Scholar Commons University of South Florida, FL.

Flahaut, B., M. Mouchart, E. San Martin, and I. Thomas. 2003. The local spatial autocorrelation and the kernel method for identifying black zones, a comparative approach. Accident Analysis & Prevention Journal 35(6): 991 -1004.

Hamann, C. J., C. Peek-Asa, C. F. Lynch, M. Ramirez, and P. Hanley 2015. Epidemiology and spatial examination of bicycle-motor vehicle crashes in Iowa, 2001–2011. Journal of Transport & Health 2(2): 178-188.

Hess P.M., A.V. Moudon, M.C. Snyder, and K. Stanilov. 1999. Site design and pedestrian travel. Transportation. Research Record Journal 1674: 9–19.

Hess, P.M., A. V. Moudon, and J. Matlick. 2004. Pedestrian safety and transit corridors. Journal of Public Transportation 7(2):73-93.

Ivan, J., C. Wang, and N.R. Bernardo. 2000. Explaining two-lane highway crash rates using land use and hourly exposure. Accident Analysis & Prevention Journal 32(6): 787-795.

Jang, K., S. H. Park, S. Kang, K. H. Song, S. Kang, and S. Chung. 2013. Evaluation of pedestrian safety: geographical identification of pedestrian crash hotspots and evaluating risk factors for injury severity. Transportation Research Board 92nd Annual Meeting (TRB Paper #13-3433), Washington, DC.

Kaplan, S., and C. Giocomo Prato. 2015. A spatial analysis of land use and network effects on frequency and severity of cyclist-motorist crashes in the Copenhagen region. Traffic Injury Prevention Journal 16(7): 724-731.

Kim, J.-K., S. Kim, G. F. Ulfarsson, and L. Porrello. 2006. Bicyclist injury severities in bicycle-motor vehicle accidents. Accident Analysis & Prevention Journal 39: 238-251.

Klop, J. and A. Khattak. 1999. Factors influencing bicycle crash severity on two-lane undivided roadways in North Carolina. Transportation Research Board 78th Annual Meeting (TRB Paper #991109), Washington, DC.

Kravetz, D. and R. Noland. 2012. A spatial analysis of income disparities in pedestrian safety in Northern New Jersey: is there an environmental justice issue? Transportation Research Board 91st Annual Meeting (TRB Paper #12-3705), Washington, DC.

LaMondia, J.J. and J. C. Duthie. 2012. Analysis of factors influencing bicycle-vehicle interactions on urban roadways by Ordered probit regression. Transportation Research Record Journal 2314: 81-88

Ma, Z., H. Zhang, R. Qiao, and Y. Yang. 2015. Modeling traffic accident frequency on a freeway using Random effect negative binomial model. ASCE Library. CICTP 2015: 3017-3026.

Martinez-Ruiz, V., O. Lardelli-Claret, E. Jimenez-Mejias, C. Amezcua-Prieto, J. Jimenez-Moleon, and J. Luna del Castilla. 2013. Risk factors for causing road crashes involving cyclists: an application of a Quasi-induced exposure method. Accident Analysis & Prevention Journal 51:228-237.

McDonald, N. 2008. Critical factors for active transportation to school among low-income and Minority students: evidence from the 2001 National household Travel Survey. American Journal of Preventive Medicine 34(4): 341-344.

McMillan, T. 2007. The relative influence of urban form on a child's travel mode to school. Transportation Research Part A, 41:69-79

Miaou, S. P. (1994). The relationship between truck accidents and geometric design of road sections: Poisson Versus Negative binomial regressions. Accident Analysis & Prevention Journal 26(4): 471-482.

Moore, D. N., W. H. Schneider IV, P. T. Savolainen, and M. Farzaneh. 2011. Mixed logit analysis of bicyclist injury severity resulting from motor vehicle crashes at intersection and non-intersection locations. Accident Analysis & Prevention Journal 43(3): 621-630.

Nashad, T., S. Yasmin, N. Eluru, J. Lee, and M. A. Abdel-Aty. 2016. Joint modeling of pedestrian and bicycle crashes: Copula-based approach. Transportation Research Record Journal 2601: 119-127.

NHTSA. Traffic Safety Facts 2013 Data. Bicyclists and other cyclists. DOT HS 812 151, May 2015.

NHTSA. Safety In Numbers Newsletter: Bicycles. Preventing two-wheeled tragedies: the mistake we all make, July 2014.

Nordback, K., W. E. Marshall, and B. N. Janson. 2014. Bicyclist safety performance functions for a U.S. city. Accident Analysis & Prevention Journal 65: 114-122.

O'Donnell, C. J. and D. H. Connor. 1996. Predicting the severity of motor vehicle accident injuries using models of Ordered multiple choice. Accident Analysis & Prevention Journal 28(6):739-753.

Poch, M. and Mannering, F.L. 1996. Negative binomial analysis of intersection accident frequencies. Journal of Transportation Engineering 122: 105-113.

Pucher, J. and L. Dijkstra. 2003. Promoting safe walking and cycling to improve public health: lessons from the Netherlands and Germany. American Journal of Public Health 93(9): 1509-1516.

Pulugurtha, S. S. and A. Nujjetty. 2012. Assessment of models to estimate crashes at intersections: with and without using traffic volume. Transportation Research Board 91st Annual Meeting (TRB Paper # 12-2880), Washington, DC.

Pulugurtha, S. S. and M. Agurla. 2012a. Geospatial methods and statistical models to estimate pedestrian activity at a bus-stop. Transportation Research Board 91st Annual Meeting (TRB Paper #12-2862), Washington, DC.

Pulugurtha, S. S. and M. Agurla. 2012b. Assessment of models to estimate bus-stop level transit ridership using spatial modeling methods. Journal of Public Transportation 15(1): 33-52.

Pulugurtha, S. S. and S. S. Nambisan. 2003. A methodology to identify high pedestrian crash locations: an illustration using the Las Vegas metro area. Transportation Research Board 82nd Annual Meeting, Washington, DC.

Pulugurtha, S. S. and E. Penkey. 2010. Assessing use of pedestrian crash data to identify unsafe transit service segments for safety improvements. Transportation Research Record Journal 2198: 93-102.

Pulugurtha, S. S. and S. R. Repaka. 2011. An assessment of models to estimate pedestrian demand based on the level of activity. Journal of Advanced Transportation 47:190-205.

Pulugurtha, S. S. and S. R. Repaka. 2008. Models to measure pedestrian activity at intersection. ASCE-Transportation Land Use Planning and Air Quality Congress 2007, Orlando, FL.

Pulugurtha, S. S., K. Krishnakumar, and S. Nambisan. 2005. Identification and ranking of high pedestrian crash zones using GIS. Annual ESRI International User Conference 2005, San Diego, California.

Pulugurtha, S. S., K. Krishnakumar, and S. Nambisan. 2007. New methods to identify and rank high pedestrian crash zones: an illustration. Accident Analysis & Prevention Journal 39: 800-811.

Pulugurtha, S. S. and S. Imran. 2013. Spatial variations in pedestrian and bicycle level-of-service (LOS) for infrastructure planning and resource allocation. ASCE Green Streets, Highways, and Development Conference Proceedings, 421-432.

Pulugurtha, S. S. and V. Vanapalli. 2008. Hazardous bus stops identification: an illustration using GIS. Journal of Public Transportation 11(2): 65-83.

Pulugurtha, S. S. and V. Thakur. 2015. Evaluating the effectiveness of on-street bicycle lane and assessing risk to bicyclists in Charlotte, North Carolina. Accident Analysis & Prevention Journal 76:34-41.

Reynolds, C., M. Harris, K. Teschke, P. Cripton, and M. Winters. 2009. The impact of transportation infrastructure on bicycling injuries and crashes: a review of the literature. Environmental Health 8: 1-47.

Roess, R., E. Prassas, and W. McShane. 2004. Traffic engineering, 3rd Edition. Pearson Prentice Hall, New Jersey, NJ.

Shankar, V., F. Mannering, and W. Barfield. 1995. Effect of roadway geometrics and environmental factors on rural freeway accident frequencies. Accident Analysis & Prevention Journal 27: 371-389.

Shankar, V., J. Milton, and F. Mannering. 1997. Modeling accident frequencies as zero-altered probability processes: an empirical inquiry. Accident Analysis & Prevention Journal 29: 829-837.

Soneji, M., G. Zhang, and S. Bajwa. 2009. Influencing road safety: critical factors during the project life cycle. 2009. Australasian Road Safety Research, Policing and Education Conference 2009(10-13):848-850.

Stipancic, J., S. Zangenehpour, and L. Miranda-Moreneo. 2015. Segmented ordered logit analysis of gender and bicycle-vehicle conflict occurrence at urban intersections. Transportation Research Board 2015 Annual Meeting (TRB Paper # 15-1339), Washington, DC.

Strauss, J., L. F. Miranda-Moreno, and P. Morency. 2013. Cyclist activity and injury risk analysis at signalized intersections: a Bayesian modelling approach. Accident Analysis & Prevention Journal. 59: 9- 17.

Vogt, A. and J. Bared. 1998. Accident models for two-lane rural segments and intersections. Transportation Research Record. 1635:18-29.

Wachtel, A. and D. Lewiston. 1994. Risk factors for bicycle-motor vehicle collisions at intersections. Institute of Transportation Engineers Journal. 1994: 30-35.

Wang, C., L. Lu, and J. Lu. 2015. Statistical analysis of bicyclists' injury severity at unsignalized intersections. Traffic Injury Prevention Journal 16(5): 507-512.
Wang, G., L. Ma, and X. Yan. 2015. Modeling traffic accident counts with a parameterization of discrete choice distributions. ASCE Library, CICTP 2015: 3254-3261.

Wang, Z., P. S. Lin, H. Chen, J. Lu, and W. Deng. 2013. Modeling impacts of access design and spatial pattern on crash risks of non-motorists on urban multilane highways in Florida. Transportation Research Board 92nd Annual Meeting (TRB Paper #13-0386), Washington, DC.

Washington, S. P., M. G. Karlaftis, and F. L. Mannering. 2003. Statistical and econometric methods for transportation data analysis. Chapman and Hall Press, Boca Raton, FL.

Wei, F. and G. Lovegrove. 2012. An empirical tool to evaluate the safety of cyclists: community based, macro-level collision prediction models using Negative binomial regression. Accident analysis & Prevention Journal 2782.05.018.

Yamada, I. and J. –C. Thill. 2004. Comparison of planar and network K-functions in traffic accident analysis. Journal of Transport Geography 12(2): 149-158.

Yasmin, S. and N. Eluru. 2016. Latent segmentation based count models: analysis of bicycle safety in Montreal and Toronto. Accident Analysis & Prevention Journal 95(Part A): 157-171.

Zahabi, S. A. H., J. Strauss, K. Manaugh, and L. F. Miranda-Moreno. 2011. Estimating potential effect of speed limits, built environment, and other factors on severity of pedestrian and cyclist injuries in crashes. Transportation Research Record Journal 2247: 81-90.

Zhang, Y., J. Bigham, Z. Li, D. Ragland, and X. Chen. 2013. Associations between road network structure and pedestrian-bicyclist accidents. Transportation Research Board 92nd Annual Meeting (TRB Paper #13-4316), Washington, DC.