GEOSPATIAL AND STATISTICAL METHODS
TO MODEL INTRACITY TRUCK CRASHES

by

NagaSwethaPasupuleti

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

2012

Approved by:

_____
Dr. Srinivas S.Pulugurtha

_____
Dr. Martin R. Kane

_____
Dr. Craig J. Allan

_____
Dr. Edwin Hauser

_____
Dr. RajaramJanardhanam

_____
Dr. Anthony Brizendine

ABSTRACT

NAGA SWETHA PASUPULETI Geospatial and statistical methods to model intracity truck crashes (Under the guidance of Dr. SRINIVAS S. PULUGURTHA)

In recent years, there has been a renewed interest in statistical ranking criteria to identify hot spots on road networks. These criteria potentially represent high crash risk zones for further engineering evaluation and safety improvement. Many studies also focused on the development of crash estimation models to quantify the safety effects of geometric, traffic, and environmental factors on expected number of total, fatal, injury, and/or property damage crashes at specific locations. However, freight safety, specifically truck safety, was meagerly addressed. Trucks and long-combination vehicles (LCVs) that carry approximately 70% freight have significant potential in triggering crash occurrences on roads, mostly severe crashes. Truck transportation is therefore attracting more and more attention due to its effect on safety and operational performance as well as rapid industrial growth.

Most of the past research on truck safety focused on intercity or interstate truck trips. Intracity truck safety related studies or research was hardly pursued. The major research objectives of this dissertation are:1) to develop a geospatial method to identify high truck crash zones, 2) to evaluate the use of different ranking methods for prioritization and allocation of resources, 3) to investigate the relations between intra city truck crash occurrences and various predictor variables (on- and off-network characteristics) to provide greater insights regarding crash occurrence and effective countermeasures, and 4)to develop truck crash prediction models.

The prioritization of high truck crash zones was performed by identifying truck crash hot spots and ranking them based on several parameters. Geospatial methods along with statistical methods were deployed to understand the relationships between geometric road conditions, land use characteristics, demographic, and socio-economic characteristics and truck crashes. Truck crash estimation models were then developed using selected on- and off- network characteristics data. To assess the suitability of these models, several goodness-of-fit statistics were computed. The geospatial methods and development of truck crash estimation models are illustrated using data for the city of Charlotte, North Carolina for the year 2008. It was found that on-off network characteristics, socio-economic characteristics and demographic characteristics that are within the 0.5-mile proximity have a vital influence on truck crash occurrence.

The findings from the research are expected to provide information and methods on identifying truck crash zones and the likelihood of a truck crash occurrence due to intracity trips and its relationship with on- and off-network characteristics of a region. Furthermore, this research is expected to aid significantly in the process of selecting meaningful countermeasures to improve safety of users on roads.

ACKNOWLEDGEMENTS

convinced me to come to the United States for higher education. Last but never the least, without the immense co-operation of my most beautiful daughter Baby Nakshatra, this uphill task of Doctorate degree would not have been possible.

DEDICATION

To

My Beloved Mother

Mrs. SaradaPasupuleti

Energy, love, motivation behind everything that I do and achieve

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1: INTRODUCTION

Freight transportation is key to economic growth. Transportation planners and engineers are not only concerned with the freight shipment, but also with their safe movement. According to the Commodity Flow Survey of 2007, trucks move 70.7% of all freight by value. The nation's 117 million households, 7.6 million business establishments, and 89,500 governmental units are part of an enormous economy that demands the efficient movement of freight (Freight Facts and FIGUREs, USDOT, 2010). Freight transportations categorized in many different ways depending on the mode, type of vehicle/equipment, and commodity. According to the Freight Facts and FIGUREs – 2009, trucks carried 11,539 millions of tons in 2002 and an estimated 13,243 millions of tons in 2008 (59.7% and 61.61% of total shipments by all modes, respectively) in the United States.

For any mode of transportation, safety is a very important aspect. Safety and mobility, be it goods or people, should go hand-in-hand to run the system efficiently. In spite of ongoing engineering, education and enforcement efforts, safety on roads is still a major issue in the transportation world. Traffic crashes are the seventh leading cause of death in the United States according to United States Department of Transportation (USDOT) and Bureau of Transportation Statistics (BTS) (2001).The global economic losses due to road traffic crashes exceed US$500 billion. As per World Health

Organization (WHO) 2009 records, over 1.2 million people are killed and over 20 million injured globally in crashes every year. In the United States alone, 37,261 people were killed in 34,017 fatal crashes, and another 2.34 million were injured in 1.63 million injury crashes during 2008 (NHTSA, 2009).

According to the Center for National Truck and Bus Safety (CNTBS), the number of persons killed in crashes involving a truck increased from 5,314 in 2002 to 5,537 in 2006, though it decreased to 4,508 in 2008. Statistics indicate that 24% of fatal truck crashes occur on highways and about 59% of large truck fatal crashes occur on undivided highways that do not have controlled access and have signalized intersections (USDOT, 2005). According to surveys, large trucks constitute approximately 4% of all the vehicles on the roads of United States. However, they are involved in 8% of all vehicles in fatal crashes per year (Zhu and Srinivasan, 2010).

The CNTBS in their 2008 annual facts book mentioned that an average of about 5,200 trucks is involved in fatal crashes each year. Tractors pulling one semitrailer are the most common truck configuration involved in crashes, accounting for about 60% of all trucks involved in fatal crashes. The report also stated that California, Florida and Texas had the greatest number of truck involvements over the period 2002 to 2008. In addition, the number of truck drivers killed nationwide in traffic crashes increased from 664 in 2002 to 784 in 2006, and later decreased to 639 in 2008. An average of 363 pedestrians and 87 cyclists (bicycles, unicycles, and tricycles) are killed each year in traffic crashes involving trucks. Deaths, injuries, and property damages due to these crashes are not only a major cause of personal suffering and financial loss to the victims, their families and friends, but also to the society. Therefore, vehicle crashes have become a major social

problem in the United States (USDOT and BTS, 2001), and how to improve traffic safety has become a major societal concern (Evans, 2004).These truck crashes impose a variety of costs on the vehicle and its driver, and other drivers either directly or indirectly.

The above facts shed light on the need for an in-depth study of the reasons for high truck crashes. Though there has been a lot of ongoing research in the field of transportation safety, truck safety, in particular, has not been explored much. It is well known and widely accepted that crashes involving trucks are generally more severe compared to other types of crashes. There have been a few studies on truck safety but most of them focused on trucks transporting between states or cities, not on intracity roads. Intracity truck crashes account for a good percentage of truck crashes that occur. Though most of them are not fatal, they impose a huge impact. (NHTSA 2006, FMCSA 2009).

This research aims to address the relatively neglected aspect of intra city truck safety. Before beginning more in-depth discussions about trucks and their safety concerns, it is important to have an introduction about what trucks are, how different they are from other automobiles, and other related basic information.

While trucks are of various sizes and volume, they are classified into various categories based on their weight, the number of axles, or tonnage carried. Based on the number and spacing of axles, they are classified into 13 different categories (Federal Highway Administration, 1985). They are:

Class 1: Motorcycles **-** All two- or three-wheeled motorized vehicles. This category includes motorcycles, motor scooters, mopeds, motor-powered bicycles, and three-wheel motorcycles.

Class 2: Passenger Cars – All sedans, coupes, and station wagons primarily for passengers and including those passenger cars pulling recreational or other light trailers.

Class 3: Other Two-Axle, Four-Tire Single Unit Vehicles – All two-axle, four-tire vehicles, excluding passenger cars. Included in this classification are pickups, panels, vans, and other vehicles such as campers, motor homes, ambulances, hearses, carryalls, and minibuses.

Class 4: Buses - All vehicles manufactured as traditional passenger-carrying buses with two axles and six tires or three or more axles.

Class 5: Two-Axle, Six-Tire, Single-Unit Trucks – All vehicles on a single frame, including trucks, camping and recreational vehicles, motor homes, etc. with two axles and dual rear wheels.

Class 6: Three-Axle Single-Unit Trucks – All vehicles on a single frame, including trucks, camping and recreational vehicles, motor homes, etc. with three axles.

Class 7: Four-or-More-Axle Single-Unit Trucks – All trucks on a single frame with four or more axles.

Class 8: Four-or-Fewer-Axle Single-Trailer Trucks – All vehicles with four or fewer axles consisting of two units, one of which is a tractor or straight truck power unit.

Class 9: Five-Axle Single-Trailer Trucks – All five-axle vehicles consisting of two units, one of which is a tractor or straight truck power unit.

Class 10: Six-or-More-Axle Single-Trailer Trucks – All vehicles with six or more axles consisting of two units, one of which is a tractor or straight truck power unit.

Class 11: Five-or-Fewer-Axle Multitrailer Trucks – All vehicles with five or fewer axles consisting of three or more units, one of which is a tractor or straight truck power unit.

Class 12: Six-Axle Multitrailer Trucks – All six-axle vehicles consisting of three or more units, one of which is a tractor or straight truck power unit.

Class 13: Seven-or-More-Axle Multitrailer Trucks – All vehicles with seven or more axles consisting of three or more units.

Overall, trucks involve a wide array of equipment, from small delivery vans and pick-up trucks to 18-wheelers. Truck equipment-type information is important in the application of models for congestion, air quality, safety, and pavement impact analyses. In this particular research, trucks with two axles and six tires or more are considered for analysis

Since the highway infrastructure is a shared-use facility, huge trucks share the same network as personal automobiles and buses, which entail high interaction with passenger vehicles and small trucks. The higher the interaction and conflicts, the higher the risk of a crash occurrence.

The Federal Motor Carrier Safety Administration (FMCSA) has recognized the need and has given high priority to research regarding collisions between large trucks (gross vehicle weight > 4,540 kilograms or 10,000 pounds) and other vehicles on the roadway (Minnesota Statewide Heavy Vehicle Safety Plan,2005).

Considering the massiveness of truck transportation and its impact on mobility, safety and economy, the current research focuses on truck-based movement, i.e., all vehicles with more than 2 axles,. Minimizing these crash costs and improving safety are major goals for many researchers and institutions.

1.1 Background

According to the Commodity Flow Survey of 2007, trucks transport 70.7% of all freight by value, while the remaining is transported by other modes or means. Furthermore, the freight transported by trucks is 68.8% of all freight by weight, and 39.8% by ton-miles (USDOT/BTS, 2008). The figures clearly explain the importance of large truck traffic. However, because of this substantial volume of truck traffic, its unique nature, its drivers along with the design/weight-related problems, trucks are one of the major reasons of crashes, injuries and fatalities on roads. (Commodity Flow Survey, 2007)

In 1998, large trucks accounted for 7% of the total vehicle miles traveled but were involved in 13% of all traffic fatalities (5,374 of 41,471) (NHTSA, 2000). In truck crashes, the car's occupants are much more likely to be killed than the truck driver (78% of the fatalities were car occupants) or injured (76% of the injuries were sustained by car occupants).

As per Federal Highway Administration (FHWA), more than 5,000 people lost their lives because of large truck crashes in the year 2005. Furthermore, 114,000 people are injured in the United States during the same year because of road crashes involving large trucks with gross vehicle weight rating greater than 10,000 pounds (NHTSA, 2006).

In 2007, 413,000 large trucks were involved in road crashes that resulted in 4,808 fatalities. This is certainly an alarming figure as it accounted for 12% of the total fatalities, while trucks accounted for about 8% of highway vehicle-miles traveled in United States (NHTSA, 2008). According to United States FMCSA's2009 report, the

number of large trucks involved in serious fatal and non-fatal crashes increased significantly by 5.9% from 2004 to 2007.

According to the 2008 National Highway Traffic Safety Administration (NHTSA) safety facts report, large trucks are much more likely to be involved in a fatal multiple-vehicle crash – as opposed to a fatal single-vehicle crash – than passenger vehicles. The large truck crash facts for 2005 report, "Trucks Involved in Fatal Accidents Fact Book – 2007" released in 2010, indicates that 5,049 fatal truck involvements were observed in 2007. Besides these, there are around 78,000 injury crashes and 341,000 property damage only crashes involving large trucks in the United States.

Decades of interdisciplinary research on vehicle crashes have revealed that there are generally five major factors affecting traffic safety and efficiency: driver behavior (about 160 million drivers in the United States), vehicle types (motorcycles to large trucks), roadway condition (design, capacity, pavement type), traffic characteristics (flow, speed, density, occupancy), and environmental factors (weather, etc.). All these factors interact with each other and influence the occurrences and severity of crashes. Driver behavior- affected by variables such as alcohol and drug use, reckless operation of vehicles, failure to properly use occupant protection devices, and fatigue - are major factors contributing to a high proportion of crashes (USDOT and BTS, 2001). Although driver maneuvers often involve considerable amount of subjective judgment, they are nonetheless the reactions to the roadway condition, traffic situation and other environmental factors. Hence, in this research roadway conditions, land use, socio-economic and demographic data are considered for in-depth study and analysis.

There is a great need for innovative methods and solutions to avoid crashes or reduce crash severity and improve safety of large truck movement. It is highly necessary in view of the revisions applied to the road or truck network to identify appropriate proactive measures and treatments to improve truck safety. To implement proper regulations to control these alarmingly increasing crash counts, the root cause of the problem need to be identified.

Regulations will have a significant effect on freight flows in a region. As an example, safety regulations such as route restrictions, truck size, and weight limitations influence routing patterns of truck movements, types of equipment used, and shipment sizes. Environmental regulations pertaining to emissions will affect equipment types, while hours of service regulations affect the time-of-day characteristics. Land use regulations may have the most significant impact on freight demand due to the inherent interrelationship between land use and transportation. For example, land use regulation on the development of warehousing facilities in a region influences truck traffic patterns and trip length distributions. To recommend such measures, it is vital to identify where truck crashes are high, when are they occurring, and what are the factors/reasons contributing to these high crash counts. These basic questions need to be answered thoroughly to come up with appropriate safety measures or techniques for appropriate allocation of resources. As on- and off-network characteristics have a bearing on the number trips, travel patterns and safety, it is important to examine the direct relationship between these characteristics and truck crashes. Hence, finding answers to where, when and why truck crashes are occurring will aid agencies to address this critical issue, and set a goal to reduce truck-involved crashes significantly. Research outputs can be informed

to land use policy and transportation planning departments for further safety enhancements.

1.2 Problem Statement

Safety can significantly influence the overall planning and design of new and rehabilitated infrastructure. Safe, Accountable, Flexible, and efficient Transportation Equity Act: A Legacy for Users (SAFETEA-LU) establishes a clear set of process and content requirements for states to establish a system that identifies hazardous locations, sections, and elements using different types of crash data. This is expected to facilitate and assist in problem identification and countermeasure selection. An analytically attractive approach is required to calculate risks in terms of crash rates based on factors of interest, using appropriate measures of exposure.

When planning new roadways, large trucks are considered more important than passenger vehicles due to certain factors such as their performance on grades, horizontal curves and different pavement types or the effect of weather on their performance. A downgrade or steep grade can not only affect truck braking performance but can also be problematic when an inattentive or inexperienced driver accelerates at an unsafe speed that may result in the loss of control of the truck. Similarly, a steep upgrade can be problematic in situations where a large truck driver is moving at a significantly slower speed compared to the rest of the vehicle traffic stream along the road section. In such situations, the drivers of other vehicles may decide to accept higher risks by passing the large trucks. Such risky initiatives may cause hazardous crashes. Therefore, a lot of attention and focus on truck movement and their potential effect is needed for safe planning aspects of transportation.

Many studies and researches focused on intersection crash estimation models (Pulugurtha and Nujjetty, 2011, 2012), pedestrian crash models (Pulugurtha and Sambhara, 2011), and so on. However, not many focused on truck crash estimation models. A few of them developed models considering network characteristics or roadway geometrics or driver characteristics, but truck crash estimation models relating crashes to demographic, socio-economic and land use characteristics were hardly found in the literature. Truck crash estimation models developed considering such characteristics would play a significant role in incorporating safety into the planning process and adopt a proactive approach to make road transportation system more effective and efficient.

Geographic Information Systems (GIS) has been proven to be a useful tool for mapping and spatial analysis in transportation studies (Miller and Shaw, 2001; Thill, 2000). Many researchers have used GIS to display crash locations on digital maps and perform various spatial analyses (including hot spot analysis) of crashes (Black, 1991; Flahaut et al., 2003; Kam, 2003; Levine et al., 1995a; Petch and Henson, 2000; Steenberghen et al., 2004). Following earlier works on network autocorrelation analysis (Black, 1992; Black and Thomas, 1998), one recent significant advancement is the network-constrained approach to conduct spatial point pattern analysis over a network (Yamada and Thill, 2004; Yamada and Thill, forthcoming). In addition, GIS enables researchers to link crash data with travel information, land use, and social-economic information to capture the relationship between crash occurrence and contributing factors.

Principles pertaining to the field of statistics have been used to develop safety performance functions of many transportation systems (Abbess et al., 1981). The most common probabilistic models used by transportation safety analysts for modeling motor vehicle crashes are the traditional Poisson and Poisson-Gamma distributions (Haleem et al., 2009). Research also documents statistical relationships based on generalized linear models (GLMs) using Poisson distribution, Poisson-Gamma distribution and Negative Binomial (NB) regression models. Other models developed by researchers in the past include Probit models (Jason, Kockelman, and Avinash, 2010), Logit models (Srinivasan, 2002), Interactive Highway Safety Design Model - IHSDM (FHWA National Highway Institute2010), and zero-inflated models (Lord et al., 2005). These studies mostly focused on pedestrian safety or crashes, in general, at intersections or mid-block.

To address truck safety issues, it is necessary to estimate the number of crashes based on the various factors that increase crash risk. In brief, the three questions that need to be answered are: where are crashes high?; when are they occurring?; and what are the causes? If these questions are addressed, it will be of great help to the transportation planners in improving safety and resource allocation

Simple crash plotting, or geo coding crash locations, cannot effectively aid in enhancing safety. Hence, this research aims to establish a geospatial method and use statistical techniques that can provide meaningful insights for truck safety studies. To fill in the gaps of previous research, this study aims to develop a methodology aiding in the identification of high-risk truck crash zones and elevate any relationship between these high-risk zones and surrounding on-/off-network characteristics. The idea behind this research is to bring in the various geometric, land use, demographic and socio-economic

characteristics into a serious purview and look into possible intracity truck safety improvements by altering them.

It is important to recognize that intracity systems are different from intercity systems. Intracity travel are based on time, cost and convenience considerations for the entire journey – from origin to destination – as against time and cost considerations for only with respect to intercity travel. Access and reliability are relatively important for intracity travel. When a crash occurs on these intracity roads especially a truck crash, it has immense impact on travel time, congestion and economy. Though there has been lot of emphasis on interstate and intercity safety studies, intracity truck safety studies are meager even today.

This dissertation aims to bridge the identified research gaps in truck safety planning process by using truck crash estimation models to answer the key questions as to where, and why truck crashes could possibly occur on intracity road networks.

1.4 Research Objectives

The primary goal of this dissertation is to research geospatial methods to prioritize zones and develop intracity truck crash estimation models for improving safety on roads. The key objectives of this dissertation are:

- To evaluate and determine where truck crashes occur and what are the causes;

- To research and develop geospatial methods to identify the locations or zones with high truck crashes;

- To evaluate ranking methods for prioritization of high truck crash zones;

- To assess the role of spatial proximity or buffer width to capture off-network characteristics data for modeling;

- To research and understand the relationship between and on- and off- network characteristics and truck crashes; and,

- To develop truck crash estimation models to help estimate intracity truck crash risk and proactively apply countermeasures.

1.3 Research Organization

The proposed methodology in this research deals with identifying high intracity truck crash zones and estimating truck crash models. Various on-network and off-network characteristics data were considered as independent variables in the development of models, along with land use characteristics.

The remainder of the dissertation consists of four chapters. Chapter II discusses various truck crash estimation/modeling techniques. A brief discussion of statistical analysis methods and distributions used in the past was also discussed in this chapter. Chapter III describes the proposed methodology, including the procedure to identify where truck crashes happen, the methods to extract data and examine the relationship between on- and off-network characteristics and truck crashes, and the geospatial and statistical methods that were adopted in developing truck crash estimation models. Chapter IV describes the selection of study area, data collection, design of sample size, and various characteristics of the data considered for the analysis. The results obtained are also discussed. A summary of findings, conclusions, and potential for future research is presented in Chapter V.

CHAPTER 2: LITERATURE REVIEW

Trucks play a vital role in freight movement, contributing to the nation's thriving economy. However, there are safety threats that accompany the mobility of trucks. It is important to identify factors that affect safety due to large trucks because it helps in developing progressive policies and regulations to enable transportation facilities to operate without compromising on efficiency and safety.

Previous research on truck crash severity has been diverse, both empirically and methodologically. From an empirical standpoint, numerous research studies have focused on the casualty of crashes and attempted to isolate the risk factors that have contributed to truck crash severity. Determinants causing truck crashes include socio-economic factors such as real gross domestic product (GDP), the unemployment rate, roadway characteristics, mileage driven, interstate highway travel, distribution of the population, age, gender, and ethnicity, traffic congestion, speed limit and weather condition.

A number of studies have attempted to identify driver characteristics (e.g., age and gender) (Campbell, 1991). From a methodological standpoint, varieties of statistical approaches were used to study crash severity. Many of these analytical methods are applied using aggregate data. The disadvantage of using aggregate data is that it can result in a loss of site- or location-specific information on the relationships between crash severity and contributing factors.

Researchers often followed two major approaches to study heavy trucks and long-combination vehicle (LCV) safety. One major approach is giving emphasis to operational characteristics and design requirements of large trucks as this can help in understanding and anticipating actual safety impacts in real conditions (Debauche and Decock, 2007; Glaeser et al., 2006; Hanley and Forkenbrock, 2005; Knight et al., 2008;Renshaw, 2007). Another approach of studying safety measures of large trucks and LCV safety evaluation is to analyze actual crash rates and their outcomes in order to identify general trends and relationships (Campbell, 1991).

Overall, a wide array of analytical methods, tools and technologies were applied to make the transportation system safer and efficient. How high crash locations were identified as well as the methods and tools used to calculate crash estimation models are discussed next.

2.1 High Crash Zones and Factors Affecting Truck Crashes

In general, research on traffic safety includes identification of dangerous locations and contributing factors based on the reported crash data (post-crash analysis). Aggregate traffic flow data and crash records were used to facilitate such research studies.

In order to organize, relate, and analyze the roadway crash data, GIS was used by many researchers. GIS analysis can help in facilitating identification of crashes and their evaluation. As an example, a GIS-based methodology was developed and used to identify high pedestrian crash zones, and rank them in decreasing order of risks involved. This ranking helped in prioritizing, identification and execution of pedestrian safety measures and treatments (Pulugurtha andNambisan, 2003).

Jovanis et al. (1983) compared the crash involvement rates for motor vehicle crashes to identify whether large trucks have higher fatality and/or injury rates than other types of vehicle. Chira-Chavala et al. (1984) applied Logit models to study the effect of truck crash severity. They found that collisions with passenger cars, collisions on dry surface roads at night, and collisions on undivided rural roads usually resulted in higher fatality and injury ratios.

Golob et al. (1987) statistically compared the mean number of injuries and fatalities by collision type and the number of involved vehicles in truck-involved freeway crashes. Alassar (1988) used a log-linear modeling approach to examine the crash severity of truck-involved crashes and identified the contributing factors (e.g., collision types and road class) for fatal and injury crashes. Saccomanno et al. (1988) used GLMs to explore the relationships between truck crash occurrence and highway geometric designs.

Campbell (1988) addressed the issue of a minimum age for drivers of large trucks by comparing the fatal crash involvement rates against driver age. Khasnabis et al. (1989) used a time series analysis to forecast truck crashes. Braver et al. (1996) examined the crash involvement rates for different truck configurations (e.g., singles and doubles) to identify if a configuration is significantly safer than others are. The effect of roadway geometry, weather, and other vital factors on the crashes involving fatal large truck-car crashes (Braver et al., 1996) established the relationship between conditions such as defiance of traffic control devices and rules, curves, slippery and roadway conditions and fatal crashes.

Duncan et al. (1998) performed a research study and examined the injury incurred on passenger-car occupants in case of rear-end collision with a heavy truck. Highway Safety Information System (HSIS) data from North Carolina for the time duration of 1993-1995was used to develop ordered-probit models. The result of their study suggested that higher speed, darkness, and grade increased the rigorousness of the injury. The results also indicated that drunk drivers and females are more prone to sustain drastically severe injuries when compared to non-drunk drivers and males, respectively. The research also strengthened the idea that snow/icy or bad weather conditions often uncharacteristically decrease the severity of injuries compared to the injuries sustained during a large-truck crash in dry and free-flow traffic conditions. The research report concluded that a car being struck in the rear led to more severe injuries when compared to a truck being struck at the rear (Duncan et al, 1998).

Chang and Mannering (1999) adopted an un-ordered discrete-choice structure to develop and model severity of injuries. The researchers divided the available data into truck-involved and non-truck-involved road crashes and explained the statistical and experimental importance of these divisions. The results emphasized that more severe injuries happen in cases of crashes involving large trucks moving at higher speeds. On the other hand, the effect of high speed is insignificant in the case of non-truck crashes. The study also looked at that the effect of turning movements (left-turn and right-turn). The results indicated that multiple-occupant vehicles in truck-involved crashes often result in considerably severe injuries.

Nukoolkit, et al. (2001) used a neural network to investigate the effect of different variables on crash severity to identify dangerous crash patterns.

In order to identify empirically significant factors contributing to truck crashes in New Jersey, a GIS-based Truck Accident Information and Management System (TAIMS) was developed. Even though the crash samples used in the study included a very short period of two years, 1998 and 1999, the results were found to be accurate and that TAIMS was adjudged as a potential tool that can help in truck crashes analysis. The results suggested that TAIMS can effectively correlate electronic maps, traffic crash data and roadway geometry data, but could not induce other useful variables such as land use information. Maintaining truck crash data and related information also proved to be time consuming and difficult to maintain consistency and completeness. This can affect the efficiency of TAIMS. Therefore, it is necessary to produce a system that can provide an appropriate database with the capability of data standardization and verification. This will help manage relevant information such as annual average daily traffic (AADT) and will help in improving the results of TAIMS (Daniel and Chin, 2003). Additionally, this will help identify and provide technologies and recommendations that may be included in planning proper safety measures against large truck crashes.

Khattak et al. (2003) used HSIS data of North Carolina for a two-year period (1996-1998) to examine the severity of injuries caused in crashes involving single-large trucks. The basic idea behind the research study was to analyze the differences between rollover and non-rollover crashes. Ordered-probit models were used to establish that rollover crashes resulted in severe injury crashes involving single-trucks.

There have been extensive investigations about the relationship between the severity of injuries in road crashes involving large trucks and the gender and age of vehicle occupants. Crashes involving large trucks, tractors or semitrailers on expressway ramps are influenced largely by the interaction between the geometry of the expressway and vehicle dynamic behavior. Underage or inexperienced drivers also affect the causation of severe injuries in road crashes involving large trucks. Based on the analysis of relationship between minimum drinking age (MDA) and highway fatalities during the period of 1975 to 1985, an11% decrease in highway fatalities involving 18-20 years old drivers can be expected if the minimum drinking age is increased from 18 to 21 years.

An increasing risk of severe injuries caused in crashes involving large trucks was reported in relationship with driver age for older and younger age driver group (Abdel-Aty, 1998). Golob and Recker (2003) stated that traffic volume plays a greater influential role on the severity of injuries during a crash than speed.

A learning process to develop a model that can be compared with a separate validation dataset to verify the accuracy was employed to make a full use of Bayesian method. The dataset used for this study was based on rural two-lane collector and arterial horizontal curves in Ohio, comprising 15,390 observations from crash records between 2002 through 2006. This study specifically stressed on the observation of impact of shoulder width, horizontal curve radius, curve length, and other traffic parameters. The results of the study suggested a significant increase in truck crashes due to both horizontal curvature and passenger vehicle volumes (Daniel and Chien, 2004).

Wang and Kockelman (2005) in their research study reported that a dramatic reduction in the probability of severe (36.5%) or fatal injuries (47.5%) for vehicle drivers who promptly use safety restraints as compared to those who often ignore safety measures.

Comparison between neural networks and Logit models was made using vehicle crash data on two-lane rural highways in Iran (Nassiri and Edrissi, 2006). A number of variables related to roadways, vehicles, environment and drivers (such as driver fatigue, head-on collision and lack of vehicle control) were found to have a significant effect on the severity of truck crashes. In addition, investigating the marginal effects of variables showed the same variables to be significant. The results of the comparison between the Logit and neural network model indicated that they both show similar patterns regarding the effects of different variables causing truck crashes, with the Logit model providing better results.

The United States Large Truck Crash Causation Study (LTCCS) (Knipling, 2008) included information about 963 crashes involving 1,241 trucks in between 2001 and 2003. Knipling (2008) used this data to compare combination truck and single-unit truck crashes. Forty-four characteristics including crash type, driver characterization, driving environment and vehicle types were studied. According to the study results, the percentage of crashes in darkness was three times higher for combination trucks in comparison to single-unit trucks.

A model that captured both signal and roadway segments was developed to observe truck safety on roadways with signalized intersections. The model can be used

for not only determining truck safety but also roadway safety (Kiattikomol, Chatterjee, Hummer, and Younger, 2008).

The relationship between driver characteristics/behavior and crash injury severity have been studied extensively. Many studies have significantly emphasized drug and/or alcohol usage by drivers and its effect on large truck crashes. The driver's alcohol or drug consumption dramatically increases the odds of having severe and even fatal crash injuries (Kim et al., 1995). Studies have suggested that drivers whose blood alcohol concentration (BAC) was greater than 0.3 gram per deciliter (g/dl)were three times more likely to be involved in crashes and incur severe fatal injuries (Bedard et al., 2002). Study performed by Khattak et al. (2002) stressed the safety issues regarding increasingly older drivers who were over the age of 65 years. This research also indicated that the alcohol consumption would increase the chances of fatal injuries in large-truck crashes. The factors that contribute to fatality in large-truck crashes aretruck drivers, vehicle defects, environmental conditions, and roadway geometry. Analyses have shown that the combination of effects of alcohol, misguided brakes, and conversation may lead to fatalities in large-truck crashes. The combination of alcohol effects with other factors including roadways geometry may increase the probability of a fatality on two-lane highways by 96%. In cases of distracted drivers, the probability of fatality may increase up to 77% when alcohol is involved in crashes on two-lane highways (Singh, 2009).

Zhu and Srinivasan (2010) used the data from the LTCCS. The empirical results obtained through ordered-probit models indicated the effects of various behaviors of drivers (variables) on the severity of injuries. Distraction of truck drivers, alcohol consumption by car drivers, and emotional factors of car drivers were taken into

consideration. Furthermore, the results also indicated the impact of other factors such as truck-driver fatigue, aggression, and seat belt usage on the severity of injuries caused in large truck crashes but it turned out to be statistically insignificant.

While most of the studies indicated a significant role played by traffic exposure on severity of injuries sustained in a crash, some studies found that the influence of AADT is almost insignificant on the probability of truck crash occurrences. Past research studies and results also indicated that AADT is highly associated with crash frequency, while there is no direct relationship between AADT and crash types.

2.2 Crash Estimation Models

In general, any research on traffic safety includes an identification of dangerous locations and contributing factors based on the recorded studies of post-crash analysis. There has been considerable research over the last three decades on the development of crash prediction or estimation models (Abbess et al., 1981; Hauer et al., 1988; Persaud andDzbik, 1993; Kulmala, 1995; Poch and Mannering, 1996; Lord, 2000; Ivan et al., 2000; Lyon et al., 2003; Miaou and Lord, 2003; Oh et al, 2003; Pulugurtha and Sambhara, 2011; Pulugurtha and Nujjetty, 2012).

Researchers, however, continuously tried to attain better methods to provide greater consistency with improved data generating mechanism, which may provide better statistical fit and suggest in-depth information, which was not available earlier. Multiple linear regression models were initially used to study the relationships between vehicle crashes and geometric conditions of road sections such as horizontal curvature, vertical grade, shoulder width and lane width (Roy Jorgensen Associates, Inc. 1978; Zegeer et al. 1987; Okamoto and Koshi 1989; Zegeer et al., 1990).

Researchers analyzing the relationship between vehicle crashes and geometric design faced problems of dealing with a huge number of road sections at random that had no reported crashes during the study period.

Maycock and Hall (1984) said that the inappropriateness of multiple linear regression models led the researchers to explore and use Poisson and NB regression models to identify factors that affect large truck safety and provide information about high crash zones. Jovanis and Chang (1986) stated that multiple linear regression models are not suitable for making probabilistic statements or guidelines for vehicle safety and crashes since most of the multiple linear regression models rely on normal assumptions and they lack the distributional property, which is necessary to explain the random and discrete vehicle crashes in an appropriate manner. As a result, most of the test statistics derived from any such multiple linear regression model are inapt and questionable (Miaou, 1993).

The Poisson and NBregression models are certainly better and more appropriate for studying and identifying safety factors affecting large truck safety when compared to multiple linear regression models because they provide a desirable distributional property that can describe vehicle crashes and geometric design relationship. However, Poisson and NBregression models are not without limitations and the relative performance of these models in establishing such relationships has not been evaluated yet (Miaou, 1993).

Hauer (2001) suggested that the major reason behind over-dispersion in a particular model is that observed units with same represented traits can have different means because of differences between unrepresented traits which have not been included

in that model. Over-dispersion may also arise from the real nature of the crash process and its impact (Mitra and Washington 2006). It can be the result of Bernoulli trial associated with unequal probability of independent events that are known as Poisson trials. Lord et al. (2005) mentioned that distributions like Poisson-gamma or Poisson-lognormal could be used for capturing over-dispersion pattern observed in crash data.

In order to enhance the traffic safety analysis, Thobias (2005) used a prediction model based on Bayesian Belief Networks technique to study the relationship among various model variables such as AADT, section length, speed limit, the number of lanes, surface width, and the number of crashes per year for each road segment. The model was designed to estimate the number of crashes per year on a roadway segment for every set of values of all of the model variables. The model incorporated GIS to display results. The result was a map of the roadway network that provided information about the estimated crash category for each roadway section. The results obtained through this newly developed model provided the estimation accuracy of 68.08% in the first set of estimations and 78% accuracy for second set.

Kim et al. (2006) studied the application of binomial multilevel models of crash types with the help of 548 motor vehicle crashes collected from 91 two-lane rural intersections in the state of Georgia. Crash estimation models estimating angle, rear-end, and sideswipe (both same direction and opposite direction) crashes were developed. This research study showed the importance of hierarchical data structure and its application in theoretical and suitable analytic approach for multilevel data, yielding intersection-related crashes by crash type.

Spatially disaggregated safety analysis models are being used in an increasing number of research efforts, which are anticipated to help in meeting the needs of region-level safety inspection. However, it is necessary to differentiate between the exposure variables and risk factors between different regions in order to provide a consistent crash frequency in different alternative studies.

Vavilikolanu (2008) performed a study in Ohio to develop appropriate estimation models that would analyze the effect of geometry, speed, and traffic volumes on large truck crashes on two-lane rural vertical curves. Three crash estimation models were developed using NBregression. These were then re-developed by using Bayesian approach so that the results obtained can be improved in the reduced models. Around 1,900 vertical curve segments with 200 large truck crashes data from 2002 to 2006 were studied. The results were evaluated to correlate vertical curve variables, which were relevant to the large truck crashes. It was found that low speed limit on crest curves and high speed limit on sag curves contributed to an increase in large truck crashes.

Huang et al.(2010)developed a Bayesian spatial model to help in accounting for county-level variations of crash risks in Florida by explicitly differentiating the variables of daily vehicle miles traveled in population. The results obtained suggested that there is no major difference in safety effects of risk factors on all crashes and severe crashes. It was also observed that the counties with higher traffic density and population density along with a higher level of urbanization had higher crash risk. The freeways are safer than arterials with respect to crash risk. Increased truck traffic volume also results in more severe crashes. The average travel time to work is negatively correlated with all types of crash risk. Population and age differences also tend to have an effect on crashes.

The young drivers tend to be involved in more crashes, whereas the increase in elderly population leads to less causality.

Park (2010) stated that while most of the studies performed for developing statistical models of crash data address a few major common properties of crash data, the real factors that are responsible for heterogeneity of crashes are still unknown to researchers. The study suggested that failure in accounting for the heterogeneity of crashes in any model could reduce the validity of the empirical results. Based on research on the different crash count models for modeling highway safety measures, the author stated that the continuous mixture of Poisson/Gamma distribution model is often preferred for accommodating over-dispersion and related issues. Yet, it is not certain that empirical frequencies of crash data will follow any particular distribution used for determining the Poisson mean rate.

A surrogate measure, Unsafe Following Condition (UFC), to estimate the probable likelihood of traffic crash with the help of individual vehicular information had been developed in the past (Park and Son, 2011). The individual vehicular information was applied to major sections of interstate highway in Virginia. Crash data and individual vehicular data were used to develop a statistical crash estimation model along with some hurdle models. The result of this research study, aimed at developing crash prediction models, showed that an aggregate UFC measure could help in predicting a traffic crash occurrence effectively. The results also showed that the Poisson model proved to be more effective than the other possible count data models in certain cases. The study proved that the aggregate UFC can work effectively to predict the probability of occurrence of rear-end crashes.

Pulugurtha and Sambhara (2011)focused on the examination of non-linear relationship between pedestrian crashes and predictor variables such as demographic characteristics (population and household units), socio-economic conditions and road network conditions which include the number of lanes, speed limit, presence of median and pedestrian and vehicular volume along with the accessibility to public transit system. The other aim of this research was to develop generalized linear pedestrian crash estimation models by the level of pedestrian activity and spatial proximity to extract site-specific data at signalized intersections. Data for 176 random signalized intersections in the city of Charlotte, North Carolina was used to examine the non-linear relationships that may be used to develop pedestrian crash estimations models. The average number of pedestrian crashes per year within 200 feet of each intersection was taken as a dependent variable, while the demographic characteristics, socio-economic conditions, land use patterns, road network characteristics and the number of transit stops were regarded as the predictor variables. To eliminate the predictor variables, which were in correlation with each other, the Pearson correlation coefficient was used. Statistical analysis suggested that increasing population, the number of transit stops, the number of approaches at an intersection and the pedestrian volume would typically result in an exponential increase in the number of pedestrian crashes. Land use predictor variables such as single-family residential area and neighborhood services have a negative effect on the pedestrian crashes in an area. This may be because of the increased level of pedestrian volume in such areas (Pulugurtha and Sambhara, 2011).

Many of the past research focused on estimating general or pedestrian crashes, while a few authors tried estimating truck crashes using driver or vehicular information. Not much literature was found on studies estimating intra city truck crashes using on- and off-network characteristics.

2.3 Limitations

Literature documents immense research to demonstrate that different crash types are related to predictor variables in different ways and that the estimation of crash models may provide more information about crash occurrence and proactively minimize occurrence of crashes through proper safety measures.

Most of the literature on truck safety focused on intercity truck crashes or total truck crashes in general. Hardly any studies are found on intracity truck crashes. While there are many crash estimation models that were developed to proactively deploy countermeasures and improve safety, very few models have been developed that can offer insights about the types of intra city truck crashes, reasons for occurring, expected crash frequency and other key numbers. Examining these details can enhance transportation safety and minimize several varieties of costs.

Interstate roadways often experience high percentage of truck trips. Despite this, only 24% of fatal truck crashes are observed on Interstate roadways (NHTSA 2006). Out of all, 59% fatal truck crashes occurred on undivided highways that do not have controlled access and have signalized intersections. The statistics suggest that truck safety research should not only concentrate on the Interstate driving conditions but should also stress on improving truck safety for secondary roadways. While many methods have been developed for segmenting roadways for safety analysis and performance functions, no

instruction is available for segmenting and modeling secondary, relatively low-volume roads.

Truck crashes are a major issue of concern because often they involve serious injuries and increase the probability of fatality. Logit modeling and neutral network modeling were the basic model structures developed to investigate factors such as vehicle, roadway, environment, and driver characteristics that may probably increase the severity of truck crashes. Researchers also tried to present models that can be used to predict the severity of truck crashes and to identify the important factors that may cause these crashes or may increase the probability of such road crashes.

In the past research, one or a few factors influencing truck crashes were considered to develop truck crash estimation models. Few addressed driver behavior. Few addressed age and gender of drivers, while a few studied the effect of road geometry or the effect of AADT on truck crashes. No past experiment or model was found addressing the effect of on- and off-network characteristics on truck crashes. Statistical models or relationships establishing the link between the various land use datasets, road geometry, driving and weather conditions and several other key aspects were not addressed.

This research aims to bridge that gap and provide insights on why, when and how large truck crashes happen on roads within a city and how they may be avoided. This lack of constructive and apt research addressing the real scenarios leading to the high truck crashes is a great motivating factor for this research.

CHAPTER 3: RESEARCH METHOD

This dissertation research aims to identify high dense truck crash zones and develop crash estimation models that explain and predict crashes involving trucks. Reliability of these models was substantiated with appropriate spatial/statistical models, thus aiding safety planners a better vision and direction in identifying high-risk truck zones and factors influencing these crash counts. Findings from this dissertation are expected to enable proper allocation of resources and funds, along with the framing and designing of necessary steps and countermeasures to promote and ensure transportation safety.

The proposed research method to achieve the goals and objectives identified in Chapter 1 includes the following steps:

1. Defining study area and data collection

2. Identifying high truck crash zones

3. Assessing spatial proximity or ideal buffer width

4. Collecting on-network characteristics and extracting off-network characteristics

5. Developing statistical models

6. Validation of models

Each of the above-mentioned steps is discussed next in detail.

3.1 Defining Study Area and Data Collection

Study area needs to be selected in such a way that it represents various levels of road classes to avoid any possible discrepancies and biased results. Traffic-related crashes are largely dependent on road functional class, speed limit, geometric conditions, driver behavior and several other environmental conditions. While other parameters have minimal role in picking a study area, choosing one with a good sample of various road functional classes is vital for this particular study. Interstates and freeways are designed for vehicles to move at higher speeds. Crashes occurring on these sections of roadways are more likely to be severe injury or fatal during off-peak hours. Likewise, major and minor arterials serve various levels of traffic and speeds and they influence the type and severity of truck crash to an extent.

Several sources provide various types of truck-related data such as tonnage, origin and destination of the commodities, and industry-wise data. However, detailed truck crash data are hardly compiled. The selection of study also depends on availability of data to support analysis.

For this research, crash data on various sections of the road functional classes were considered. Since this research focuses on intracity, truck crash trends on different road classes within the city were selected to develop truck crash estimation models.

In general, crashes largely depend on demographic, socio-economic, land use and on-network characteristics. An increase in demographic and socio-economic characteristics such as population and employment of location/zone may increase truck volume and crashes due to increased trips generated, attracted or both. Commercial and industrial spaces typically attract more truck volumes. Higher speed limit, traffic volume, and other parameters (lane width, the number of lanes, grade, sight distance, etc.)

increase crash risk. Therefore, there is a need to consider these datasets in assessing the relationship between crashes and on- and off-network characteristics, and in developing truck crash estimation models.

The selection of study area also depends on the availability of data such as:crash data, demographic characteristics (population, household units, size and type, ethnicity, age, and gender), land use characteristics (residential, commercial/retail, industrial etc.) and on-network characteristics (functional class, area type, numbers of lanes, and speed limit). They are identified and used in this research.

3.2 Identifying High Truck Crash Zones

After selecting the appropriate study area and data needed for the research, it is important to identify high truck crash zones in the study region before moving to further steps of the formulated methodology. For this purpose, several tools available in off-the-shelf GIS software were used.

A systematic approach was adopted to identify these hot spots. This approach includes:

1. Geocoding crash data

2. Applying Kernel density method

3. Ranking of high truck crash zones

3.2.1 Geocoding Crash Data

There are several ways to locate crashes onto digital maps. Crashes can be directly added if the exact geographic references of crash locations, like coordinates, are available. Address geocoding can be conducted when the exact address (e.g., street name and number, city, state, zip code) is available. In this research, linear referencing was

performed to locate crashes since the roads on which crashes occurred and their positions relative to the starting points of the routes were known.

Geocoding is the process of assigning a location, usually in the form of coordinate values to an address by comparing the descriptive location to the reference material. With geocoded addresses, address locations are spatially displayed which helps to recognize patterns within the information. In this step, the crash database with location details is imported into a GIS environment. The truck crashes are then geocoded on to the road network map. Once the required crash data is geocoded, the next step of the research was the identification of high truck crash zones.

3.2.2 Applying Kernel Density Method

Kernel density is a tool in GIS, which calculates density of chosen features in an area. High dense locations are where certain features are concentrated. In this step, high truck crash zones are mapped using Kernel density estimation method, aiding to understand where truck crashes are highly concentrated in the study area. Figure 1 depicts an example Kernel density map. Dark red spots represent high dense areas (high number of truck crashes); while light green spots are low dense areas (fewer number of truck crashes).
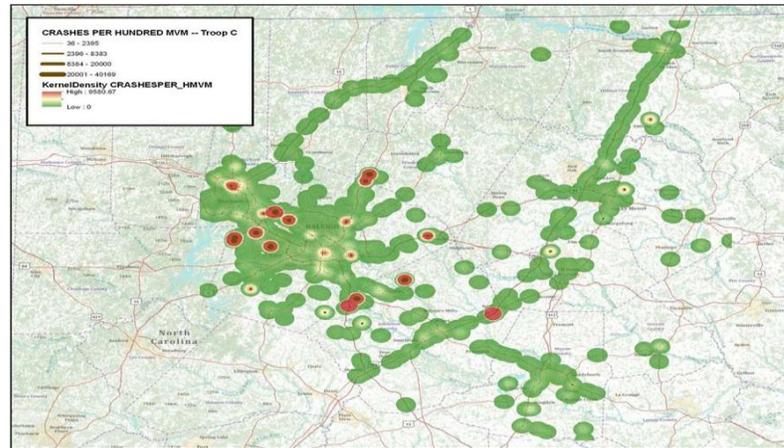
FIGURE 1: Sample Kernel density map

3.2.3 Ranking of High Truck Crash Zones

This subtask focuses on ranking high truck crash zones for allocation of resources. Identifying and ranking high crash zones plays a key role in developing efficient and effective strategies to enhance truck safety. One of the main objectives of this research is to identify "high truck crash locations" in order to allocate resources including Federal Safety Funds, for safety improvements. The criteria will help in the development of an appropriate truck safety program assisting the traffic safety managers in better understanding and identifying appropriate operating strategies to enhance truck safety. Ranking of the crash zones can also assist federal and state agencies in prioritizing funds and resources. This methodology can be implemented in any region to gain a better understanding of how truck crash locations are distributed. Even the transportation planners can take necessary steps in the regions that are highly ranked to ensure better safety norms.

Three different individual ranking methods (crash frequency, crash density and crash severity) were considered and evaluated in this research.

Crash frequency(CF): After calculating density using Kernel density tool, high truck crash zones are assigned ranks based on the truck crash count. The higher the number of truck crashes, the lower the rank that shall be assigned. For instance, a zone with most number of crashes shall be given rank 1, and so on. Therefore, all the zones are ranked based on the number of crashes that occurred.

Crash density(CD): Based on the number of truck crashes within a given area, crash density is calculated. The formula used to calculate crash density is:

Crash Density = <u>Number of Truck Crashes</u>

Area of the zone

Hence, crash density is calculated and high truck crash zones shall now be assigned ranks similar to the crash frequency method.

Crash severity(CS): Based on the crash severity and pre-established weight factors, scores are computed to rank high truck crash zones. As an example, fatal crashes are multiplied by a factor of 64, Type A, B injury crashes are multiplied by 19 and 1 multiplies Type C, and non-injury type crashes.

After converting all the type of truck crashes into a common numerical, a sum of the ranks was obtained by summing up the ranks of all the three different methods. The average of the combined rankings from crash frequency, crash density and crash severity was calculated and zones were ranked similarly as in the previous methods – the higher the sum of the ranks, the lower the rank. The combined method (called sum of the ranks

methods) combines all the three parameters helping in identifying high truck crash zones based on severity, frequency and density.

The combined method rules out any impartial or biased categorization of high truck crash zones due to merits or demerits of individual ranking methods (Pulugurtha et al., 2007). The spatial pattern of crashes analyzed in this step pinpoints statistically high dense truck crash areas and is particularly helpful in resource allocations.

3.3 Assessing Spatial Proximity or Ideal Buffer Width

The idea behind developing models based on spatial proximity is to identify the ideal buffer width to extract spatial data.

In a typical urban setting, land developments takes place along the transportation corridors. Various off-network characteristics data such as demographic, socio-economic and land use data influence traffic crashes on various adjacent roads. To quantitatively assess the influence of such characteristics, they should be captured. While it is evident that off-network characteristics data need to be captured, there is no documented evidence on the distance at which the data has to be captured. Since, this research focuses on intracity truck crashes, most of the trips would be delivery or pick up trips than long distance trips. Moreover, to capture data influencing crashes at an intra-city level, close proximity buffers are preferred. Hence, buffer widths of 0.25-, 0.5-, and 1-mile are generated around each selected high truck crash zone to extract off-network characteristics data.

3.4 Collecting On-network Characteristics and Extracting Off-network Characteristics

The focus of this step is to examine any existing relationships between the high truck crash zones and surrounding on- and off-network characteristics. This would help

identify and understand specific reasons (causes) influencing truck crash counts, which is very vital in working towards developing countermeasures to minimize truck crash counts.

In this step, features available in GIS were used to conduct a spatial analysis to collect data or extract data from the spatial data layers. Aerial photographs and field visits as well as network files maintained by local agencies were used to extract on-network characteristics.

3.4.1. Demographic and Socio-economic Characteristics

To extract site-specific demographic and socio-economic characteristics, planning variables data was overlaid on 0.25-, 0.5- and 1-mile buffers generated around each high truck crash zone. The population, the number of household units, and total employment are calculated for each high crash zone. Spatial attribute data inside these buffers were generated for modeling and analysis.

3.4.2. Land Use Characteristics

To extract specific land use characteristics and use them in truck crash estimation modeling, land use coverage was overlaid onto 0.25-, 0.5-, and 1.0-mile buffers generated around each high truck crash zone. The areas of each land use type within the buffers around each selected high truck crash zone were extracted to understand and identify the land uses that could have a significant impact on truck crashes.

3.5 Develop StatisticalModels

This step focuses on developing truck crash estimation models that provide insights on the degree to which one or more variables potentially promote positive or negative effect on truck crashes. First, variables that are not correlated to each other are

identified by testing multicollinearity between variables. In general, two variables are said to be perfectly multicollinear if the correlation between two independent variables is equal to 1 or -1, which is studied using Pearson correlation coefficient. Pearson correlation coefficient is the ratio of covariance to the product of standard deviations. It represents the strength of association or linear relationship between two variables. The higher the absolute value, stronger is the relationship. Highly correlated predictor (independent) variables could produce significant overall P-values even when the variables do not produce an effect on the dependent variables. This problem of unreliable results due to correlation among the independent variables is called as multicollinearity effect. Hence, it is recommended that the correlations among the (independent) variables considered for developing the models be in the range (-0.3, +0.3). A rational approach was adopted in identifying and eliminating variables with inter-correlations. Models were then developed as a function of land use characteristics and socio–economic and demographic characteristics. Choosing an appropriate statistical model for estimating truck crashes is a challenging task. Extensive studies have been reported on the relationship between crash characteristics (rate, frequency, fatality, injury, duration, severity, etc.) and related variables, such as weather conditions, geometric design of roads, traffic volume, road density, and driver behaviors using a variety of statistical modeling techniques. Miaou and Lum (1993) found that conventional linear regression models were not appropriate for modeling vehicle crash events on roadways. Since crashes are counts, generalized linear models (GLMs) based on Poisson, NBor Gamma distribution may be more appropriate for modeling.

GLMs are a class of statistical models that are obtained by a natural generalization of standard linear models (McCullough and Nelder, 1989). Often categorized as extensions to the standard linear regression, GLMs can incorporate various outcomes such as count, binary, proportions and positive valued continuous variables (Hilbe, 1994). This is one of the reasons for increased use of GLMs in statistics. Moreover, the normality and constant variance of errors are no longer required (McCullough and Nelder, 1989). Count regression models are best suited for analyzing skewed (non-normal) data with non-linear relationships (Elhai et al., 2008). The Poisson, NB, and other related models collectively come under GLMs.

Wood (2005) stated that GLMs have gathered recognition in recent years as useful tools for relating the number of crashes of a specified type, to explanatory variables such as traffic volume. The Poisson model is the basic and simplest count regression model (Long, 1997;Elhai et al., 2008). Unlike the normal linear regression models in which the dependent variable is transformed, in a Poisson model the probability of counts are determined by a Poisson distribution in which the coefficients are exponentiated. The Poisson log linear relationship holds true when the mean is equal to the variance. If this assumption is not valid, the standard errors, usually estimated by the maximum likelihood (ML) method, are biased and test statistics derived from the model will be incorrect. In case of over dispersed sample, NBdistribution may be used instead of a Poisson distribution.

Shankar et al. (1998) indicated that the NBmodel may be more appropriate as geometric and traffic variables are likely to have location-specific effects. The NBmodel

assumes that unobserved crash heterogeneity across sites (intersections, road segments, etc.) is Gamma distributed (Washington et al., 2003;Mitra et al. 2007).

Based on the above discussion, GLMsare more apt in developing crash estimation models in this research. Various distributions (linear, Poisson, NB, and Gamma) are considered to identify the best distribution to estimate truck crashes. These models are expected to help transportation planners in reorganization or restructuring of on- and off-network characteristics that influence truck crash counts.

Statistical models were developed using statistical software SPSS®. SPSS is a comprehensive system for analyzing data. Data can be input and used to generate reports, charts, and plots of distributions and trends. Using SPSS, descriptive statistics and complex statistical analysis can be performed easily.

Non-linear count models were developed using the independent variables that were selected from the correlation matrices. Independent variables were screened based on their significance levels (P-values).Significance level is defined as maximum allowable probability of committing a statistical error. In this dissertation, a significance level less than 0.05 were considered. Thus, independent variables with a P-value greater than or equal to 0.05 were eliminated and the analysis was repeated. Wald confidence interval was also computed, which should be typically larger (greater than 1.0). The coefficient for selected independent variables in the model will normally lie within the lower and upper limits of Wald confidence interval. Those variables whose coefficients do not lie within the Wald confidence intervals were eliminated from further analysis. Using the above process, final models were developed to estimate high truck crash zones with various data sets.

Once the models are developed, best models are picked by examining the goodness of fit statistics. For this purpose, SPSS® was used to calculate the Deviance, Scaled Deviance, Pearson Chi-Square, Scaled Pearson Chi-Square, Log Likelihood, Akaike Information Criterion (AIC) and Second Order Information Criterion (AICC) statistics for the developed models.

Deviance is defined as two times the difference of the log-likelihood for the maximum achievable model and the log likelihood under the fitted model. Pearson Chi-Square is defined as the squared difference between the observed and predicted values divided by the variance of the predicted value summed over all observations in the model. The deviance value/degrees of freedom is generally preferred to be around one. When this value is around one, the model becomes optimally dispersed i.e., neither under dispersed nor over dispersed.

The AIC is defined as $AIC = 2k - 2\ln(L)$ where k is the number of parameters in the statistical model, and L is the maximized value of the likelihood function for the estimated model. From the given set of models for the data, the preferred model is the one with the minimum AIC value. The AICC statistic helps the user to decide on the best subset of model predictors for a particular correlation structure. It is designed to evaluate which predictors best explain the response within the model framework. The statistic takes into account sample size by increasing the relative penalty for model complexity with small datasets. $AICC = -2(\ln(\text{likelihood})) + 2K*(n/(n-K-1))$ where n is the sample size. The AICC values for each scenario and buffer width were examined. Evaluation and selection of models was primarily based on AICC (lower the better).

3.6 Validation of Models

The models selected after testing for goodness of fit were considered for validation. Data for eight zones not used in model development were used to validate the developed models. The values for the independent variables are substituted into the model equation and compared to the actual crash counts. The percent difference in values of observed and predicted values for the eight zones are then calculated. The root-mean-square deviation (RMSD) or root-mean-square error (RMSE) which is frequently used to measure the differences between values predicted by a model and the actual values. RMSE is a good measure of accuracy. It is just the square root of the mean square error. Root mean square of the percentage difference is calculated in this research to observe the mean square error between the actual and calculated crash counts.

To summarize the proposed methodology, using the truck crash data gathered from various sources, high truck crash zones are identified and ranked based on various parameters. Then data layers such as land use, demographic and socio-economic characteristics were overlaid on the buffers generated around each high truck crash zones. The attributes responsible for high truck crashes are extracted using features available in a GIS.Multicollinearity between selected variables was tested. Using the final variables selected, statistical models to estimate truck crashes were then developed. Using the goodness of fit test, suitable models are selected and validated.

CHAPTER 4: DATA PROCESSING AND ANALYSIS

This chapter discusses the study area, selection of locations and various datasets to develop models, along with the analysis and results obtained.

4.1 Study Area

The study area for this research is the city ofCharlotte, North Carolina with a population of close to 700,000 (US Census Estimates, 2008). Charlotte ranks top in terms of population, traffic congestion and urbanization in the state of North Carolina. According to the 2008 United States Census population estimates, it is one of the fastest growing cities and among the twenty most populous cities in the United States. The city's downtown attracts traffic from within the Mecklenburg County and several adjacent counties. A road network map of study area is shown in Figure 2.
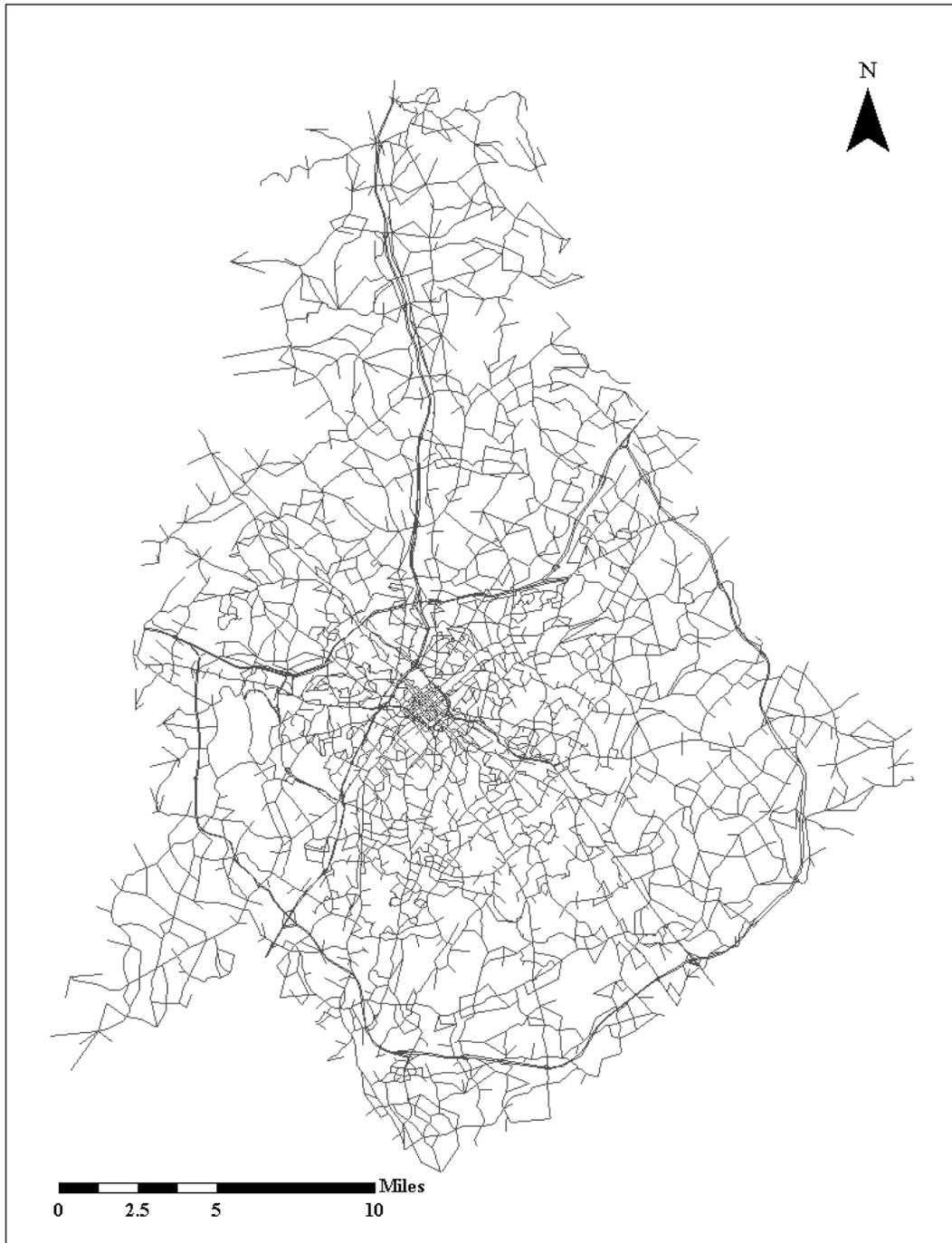
FIGURE 2: Charlotte, NCRoad Network

4.2 Data Collection

The data collection process plays a vital role in obtaining good results based on analysis and in deriving meaningful conclusions. In order to develop models to predict the occurrence of crashes, detailed information of crashes, traffic volume, the number of lanes, speed limit, other road geometry data, land use, demographic, socio-economic and network characteristics are required. Since the study area for this research is Charlotte, the highway network and traffic analysis zone (TAZ) level planning variables data are obtained from the city of Charlotte Department of Transportation (CDOT). For consistency, all the traffic, network, demographic, socio-economic and land use data for the year 2008 are used. Crash data and network characteristics such as the number of lanes, and speed limit of the study links were obtained from CDOT and the CNBTS.

4.2.1 Crash Data

Crash data is critical in developing crash estimation models. In this research, crash statistics were collected from both CDOT and CNBTS for year 2008. Crash data collected includes the harmful event, severity of crash, type of crash, and the time of crash. Figure3shows the spatial distribution of truck crashes in the study area for the year 2008.
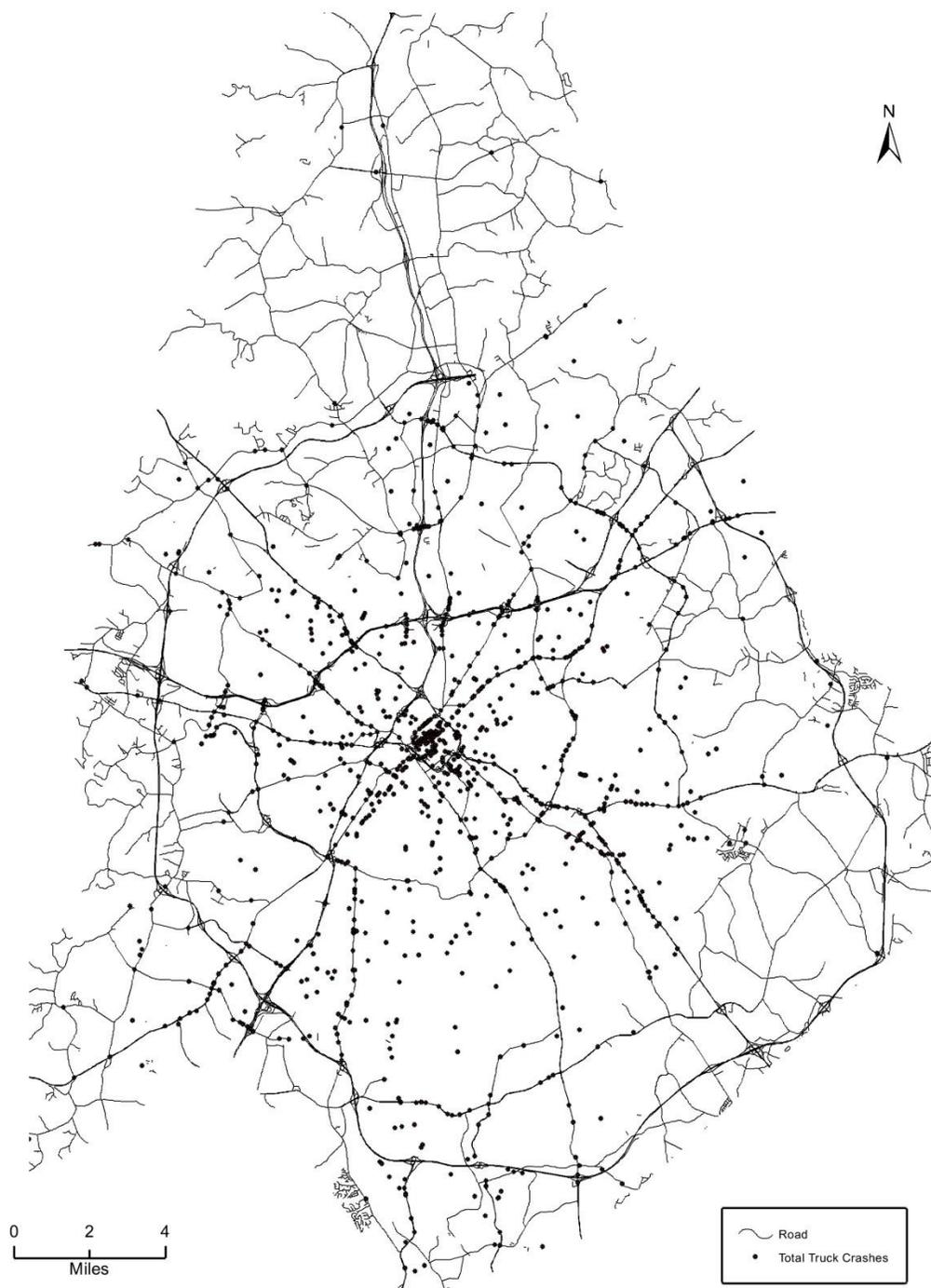
FIGURE3: Distribution of Truck Crash Data 2008

4.2.2 Socio-economic and Demographic Data

Demographic characteristics such as the number of households, population, household population, and socio-economic characteristics such as mean income and unemployment rate were considered in the analysis.

4.2.3 Land Use Data

Land use zoning characteristics such as single family and multi-family, urban residential, office, business, institutional, industrial, commercial/retail and other land uses are expected to be significant attractors and generators of truck trips. Residential land uses considered include single-family housing, multifamily housing, urban residential and primarily residential mixed-use land uses. Other land uses considered include industrial, business, commercial center, office, institutional, research district and other mixed-use districts.

4.2.4 Roadway Network Data

Crashes depend on network characteristics largely. In this research, the network characteristics for the study locations were collected fromCNBTS and CDOT. The road network characteristics like traffic volume, link or segment length, speed limit, the number of lanes, horizontal curvature, median, and the number of driveways were considered along with land use and socio-economic data for a thorough understanding of the relationship between high truck crash zones and on- and off-network characteristics.

4.3 Analysis

This part of the chapter explains the analysis performed using the methodology discussed in Chapter 3. Data was obtained for the city of Charlotte, North Carolina. The

capabilities and features available in commercial GIS software were explored along with other statistical software.

4.3.1 Descriptive Analysis

The aim of descriptive analysis is to summarize the dataset and have a better understanding of when, where, why and how truck related crashes are occurring in the study area. It involves finding answers to questions such as: What does the data consist of? How many fatal crashes occurred? How many were injury crashes? How many crashes occurred due to bad weather? How many crashes were sideswipe collisions?

The total numbers of truck crashes recorded for the year 2008 are1,280.Of which 999 crashes i.e., 78% of the total crashes are non-injury type or property damage only crashes. While 213 crashes are C type injury crashes, 54 crashes are B type injury

crashes. Fatal crashes account for hardly less than one percent. The crash distribution chart (Figure4) gives detailed breakdown of these crashes.

Figure 5 shows type of truck collisions distributions that took place in the year 2008. Of the 1,280 crashes, 427 were sideswipe collisions, 363 rear end crashes, 114 backing up crashes, 86 angle collisions and 72 left turn crashes. While 8 were head on collisions, 35 were right turn crashes and 15 were pedestrian crashes. Of the rest, 19 were roll over and 24 were due to hitting a moving object.

Figure6 shows the type of vehicles involved in the truck crashes. Single Unit Trucks involved are44% of the total truck crashes, while single unit trucks with 3 axles were involved in 25%.Semi-trailersare involved in 12% of total truck crashes while heavy trucks are involved in 14% of the total truck crashes. Rest of the crashes involved recreational vehicles and tractors.

Figure7provides an overview of the climatic conditions that prevailed during the truck crashes. Twenty percent of truck crashes occurred under cloudy conditions while 8% of truck crashes occurred under rainy conditions. The remaining 72% of the truck crashes happened when the weather was quite clear.

Crash Severity Distribution(2008)

1% 0% 0% 4%

17%

78%

- Killed
- A type injury (disabling)
- B type injury (evident)
- C type injury (possible injury)
- PDO (No injury)
- Unknown
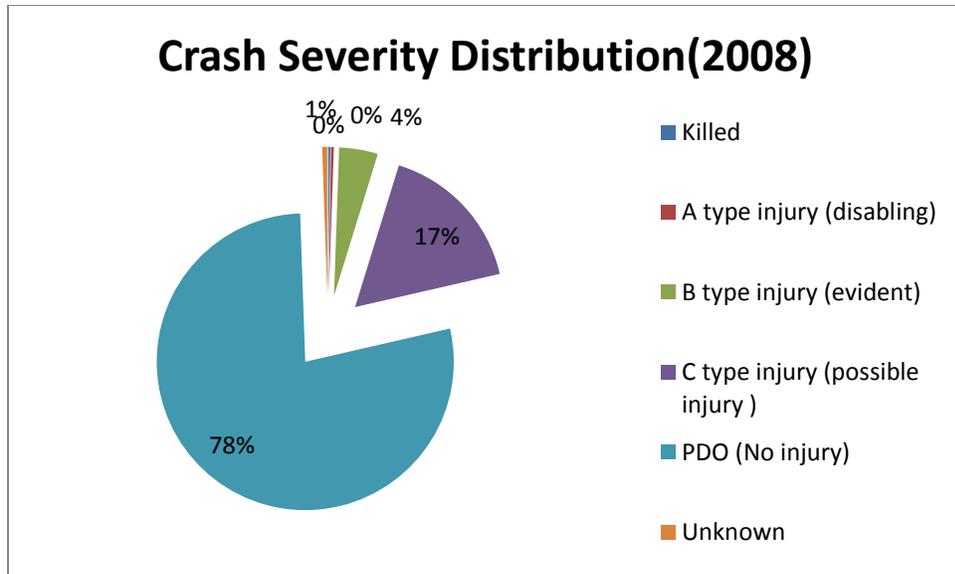
FIGURE4: Crash Severity Distribution

Collision Type Distribution (2008)

1% 3% 1%

7%

9%

5%

1%

6%

2%

28%

3%

1%

33%

- Angle
- Backing Up
- Fixed Object
- Head On
- Left turn
- Movable Object
- Rear end
- Right turn
- Ran off rd
- SideSwipe
- Roll over
- Others
- Pedestrain
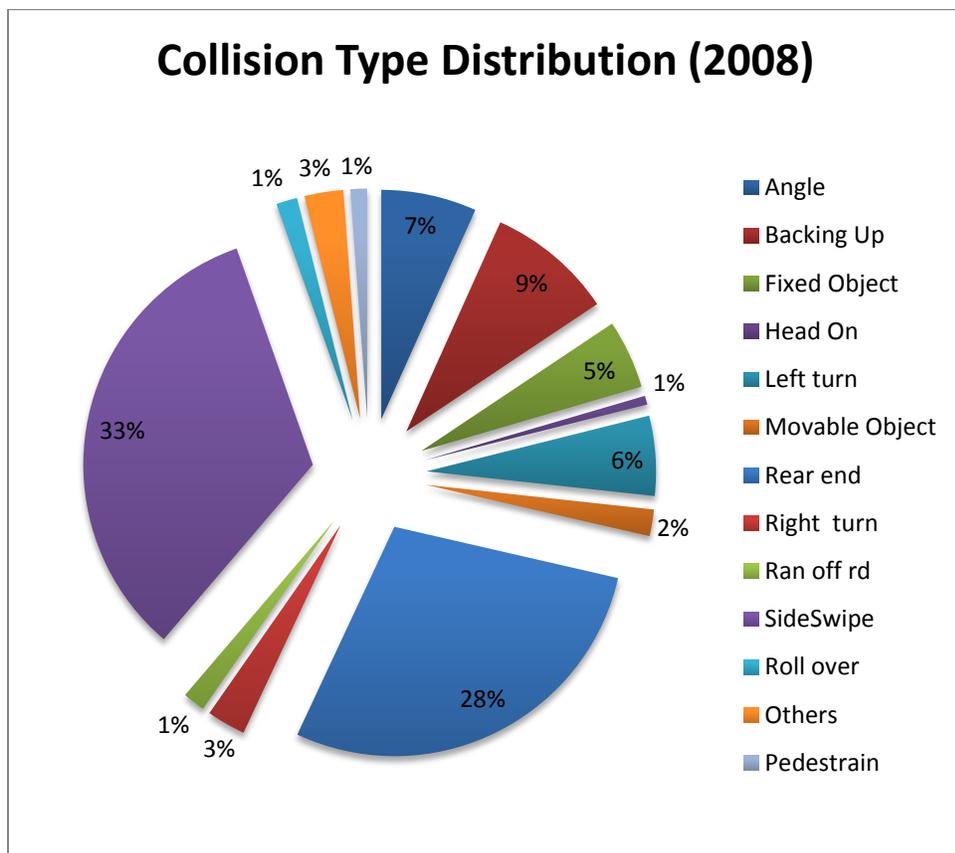
FIGURE5: Collision Type Distribution

Vehicular Classification

■	Recreational
■	single unit trucks (2 axles - 6 tire)
■	single unit trucks (3 or more axles)
■	Tractors
■	Semitrailer

1% 3% 1%
14%
12%
0%
44%
25%
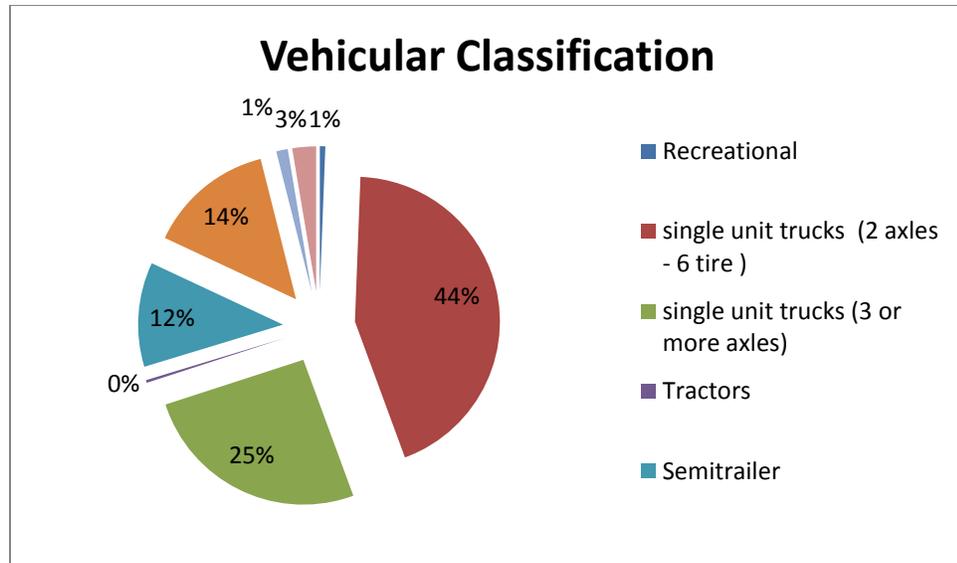
FIGURE6: Type of Vehicles Distribution

Climate Contribution

■ Fog ■ Clear ■ Cloudy ■ Other ■ Rain ■ Sleet/Hail

0%
0%
0%
8%
20%
72%
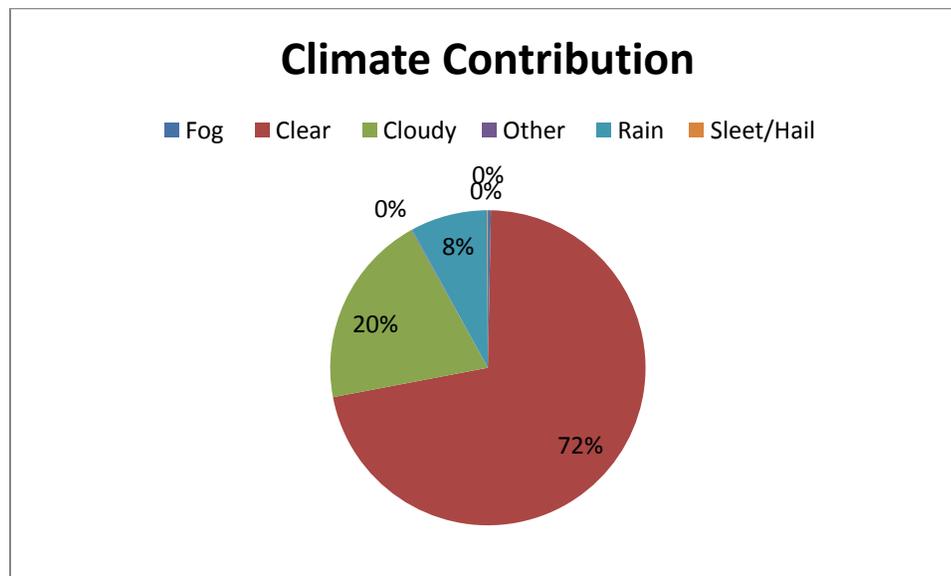
FIGURE7: Climate Distribution

4.3.2 Identifying High Truck Crash Zones

The first step of the described methodology is identifying high truck crash zones. This was done using the Kernel density method. Different Kernel density maps were generated using cell size (100 ft and 500ft), radius (330ft, 660ft, 1,320ft and 2,640 ft) and using standard deviation for classification.

Cell size is the width or height of the output, whichever is smaller divided by 250.The GIS Kernel density method considers the default size for computations. GIS measured the output extent from left to right (the shorter distance) and this came up to 126,750 ft (about 24 miles) which can be seen on scale bar. Therefore, 126,750ft divided by 250 is a cell size of 507 ft. With a 507ft - cell size, the map looked over-dispersed, highly pixilated and was tough to identify high dense locations. So based on a trial-and-error method, the cell size was manually adjusted to 100ft for better display.

The 330ft, 660ft or 2,640ftradii were found to be either too small or too large to make any meaningful interpretation. Hence, 1,320ftwas selected to identify high dense truck crash locations. Figures 8to 11 show selected Kernel density maps generated in this research. Figure8 shows the Kernel density distribution for the study area with 330 ft radius and 100ft cell size. Figure9 shows the Kernel distribution with 660 ft radius, and similarly figures 10 and 11 shows figures with 1,320 ft and 2,640ft.
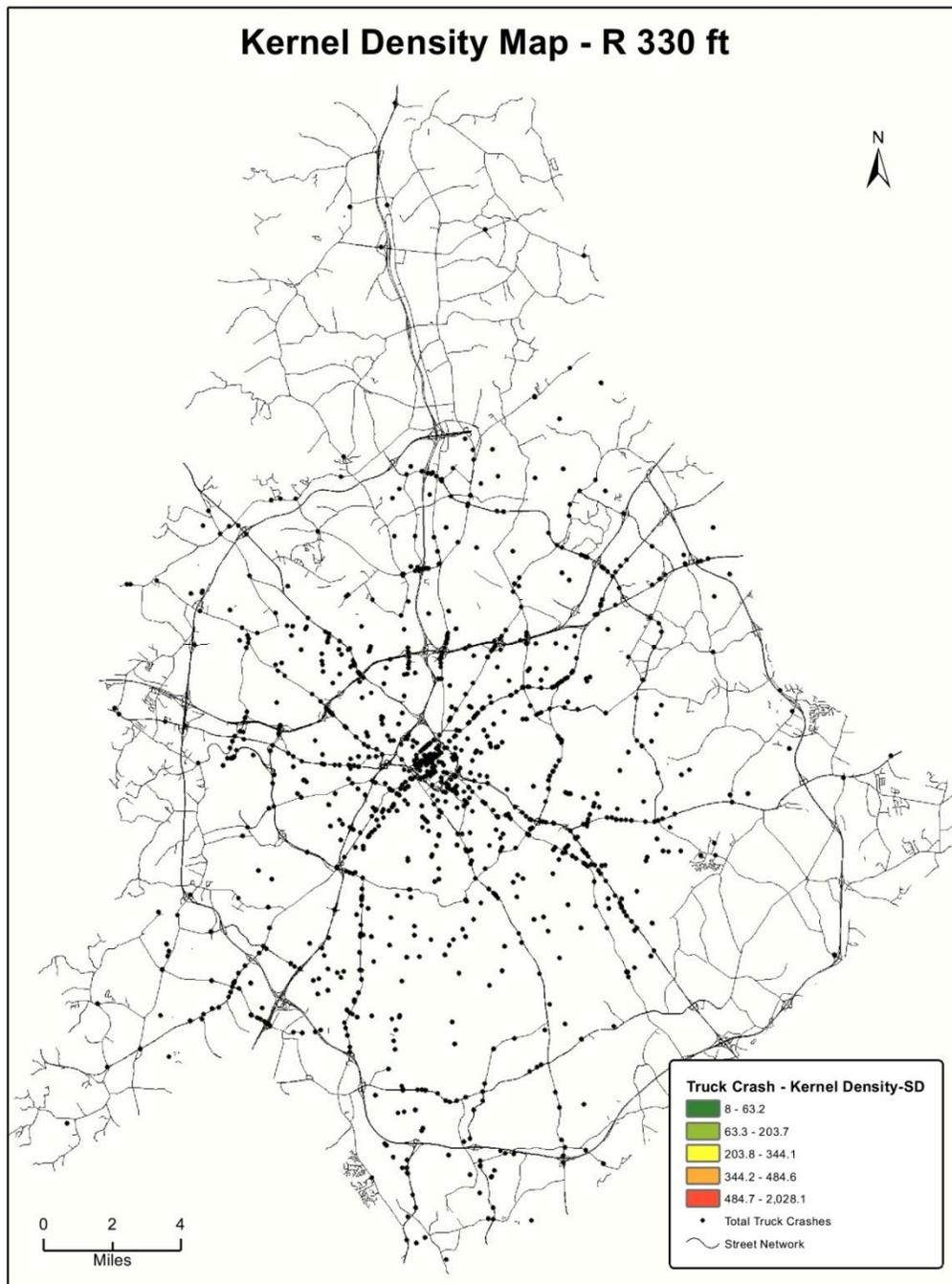
FIGURE8: Kernel Density Map - 330 ft
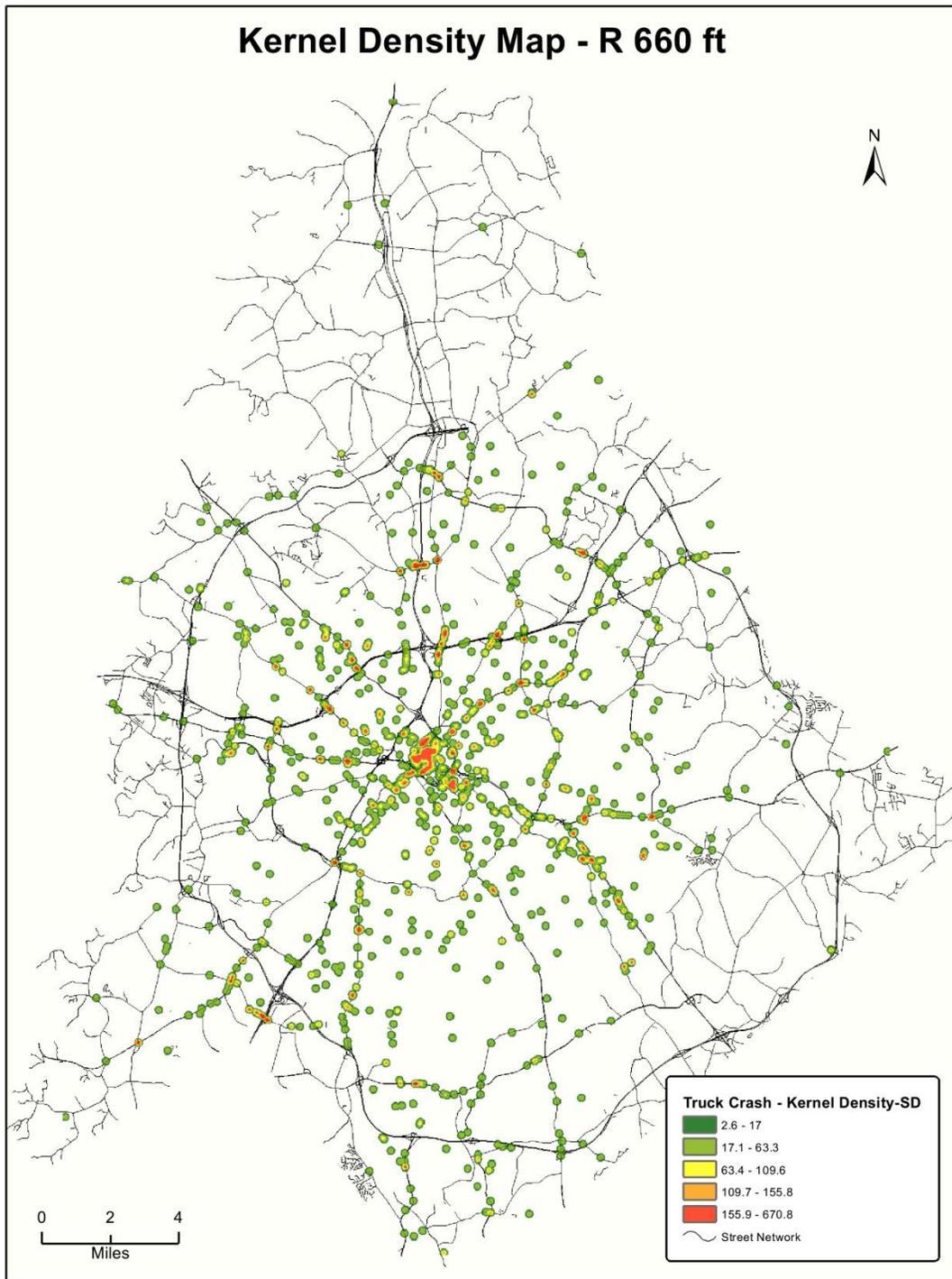
FIGURE9: Kernel Density Map - 660 ft
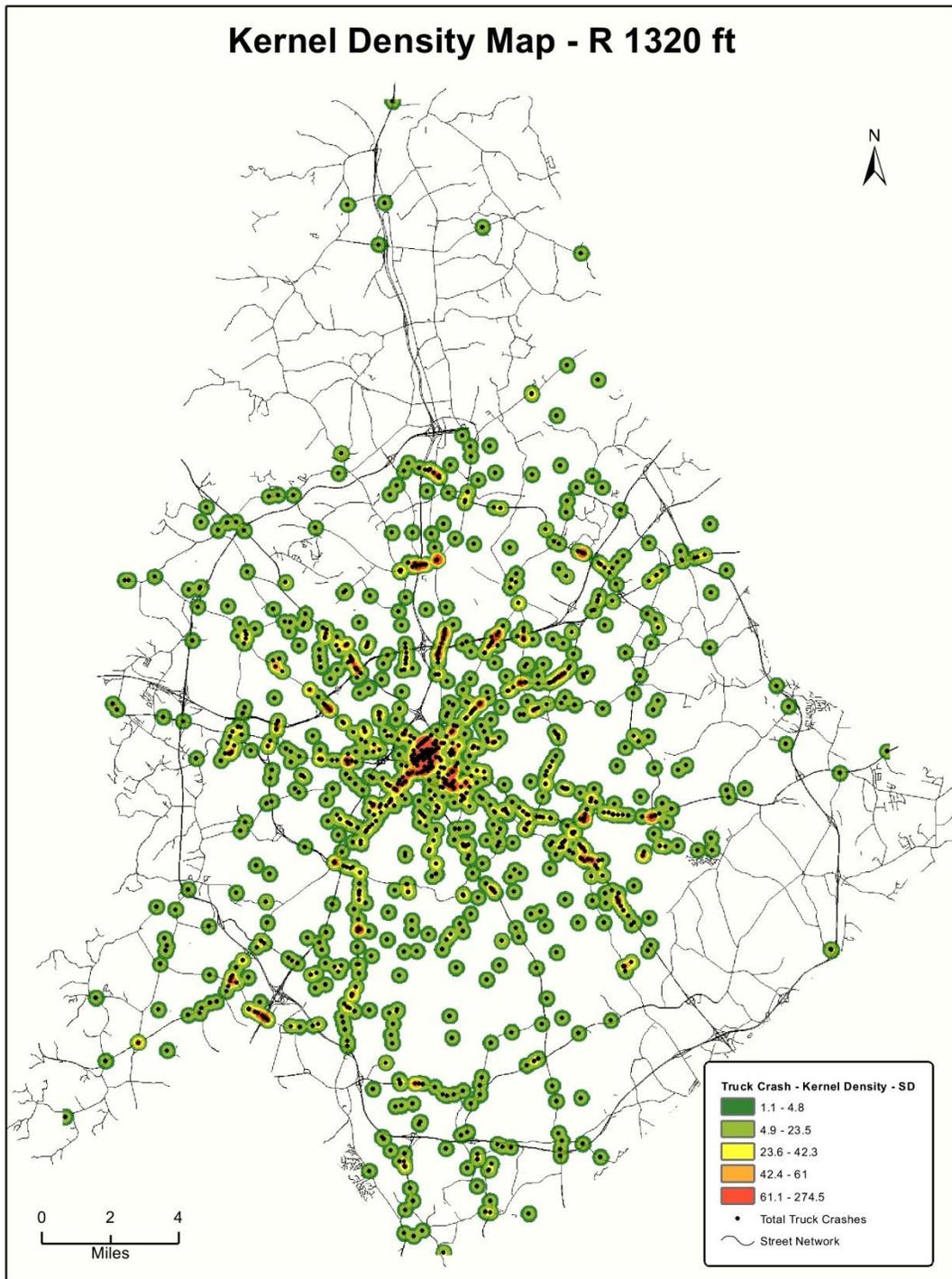
FIGURE10: Kernel Density Map - 1320 ft
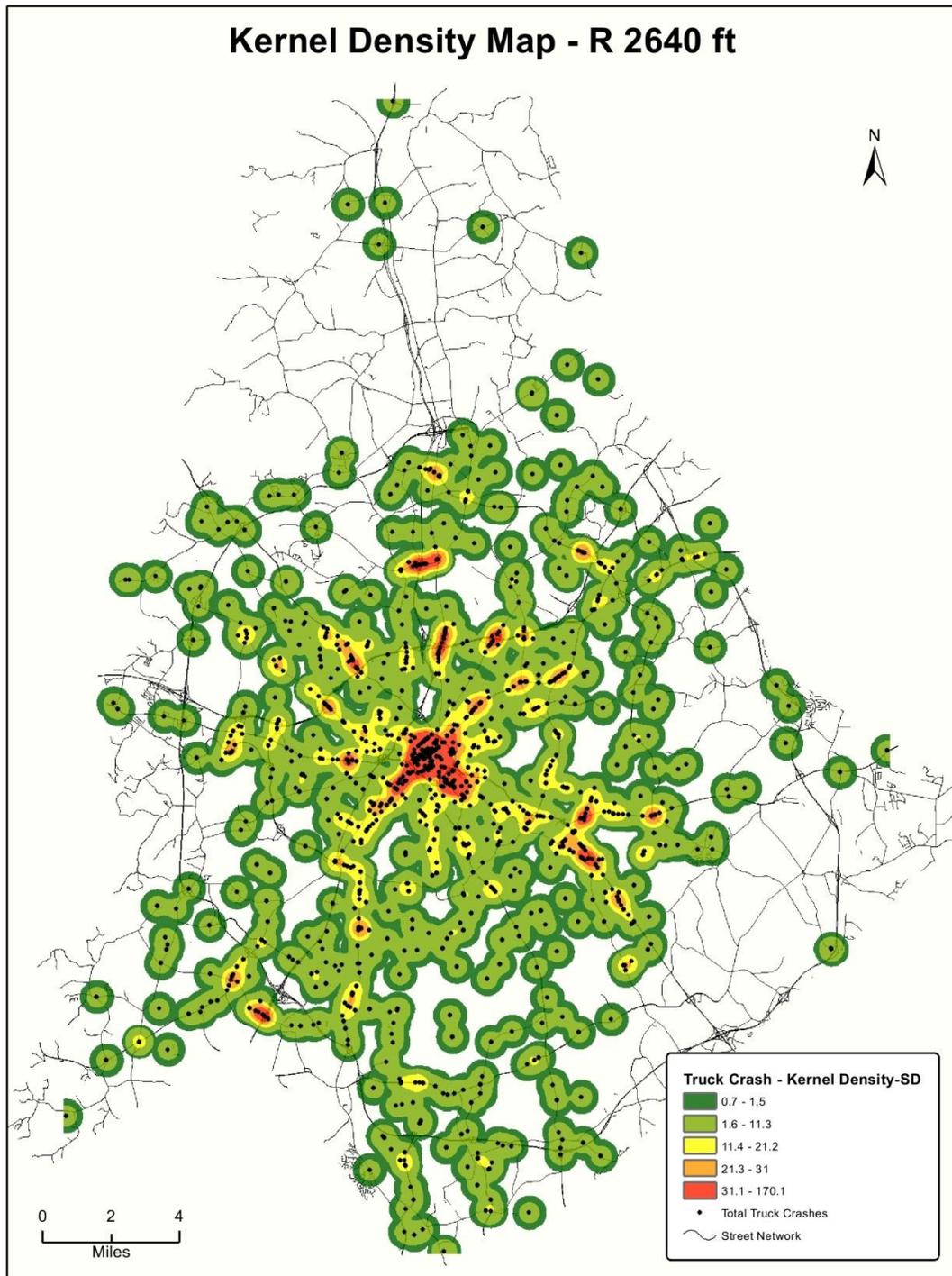
FIGURE11: Kernel Density Map - 2640 ft

Once the suitable Kernel density map is chosen, the next step in the analysis is to rank these high dense truck crash zones based on various parameters.

4.3.3 Ranking High Truck Crash Zones

Sixty linear zones were selected from the high dense crash areas using GIS and Microsoft office applications: 20 from high dense areas, 20 from medium dense, and 20 from low dense crash areas. Figure 12 shows the selected 60 zones.
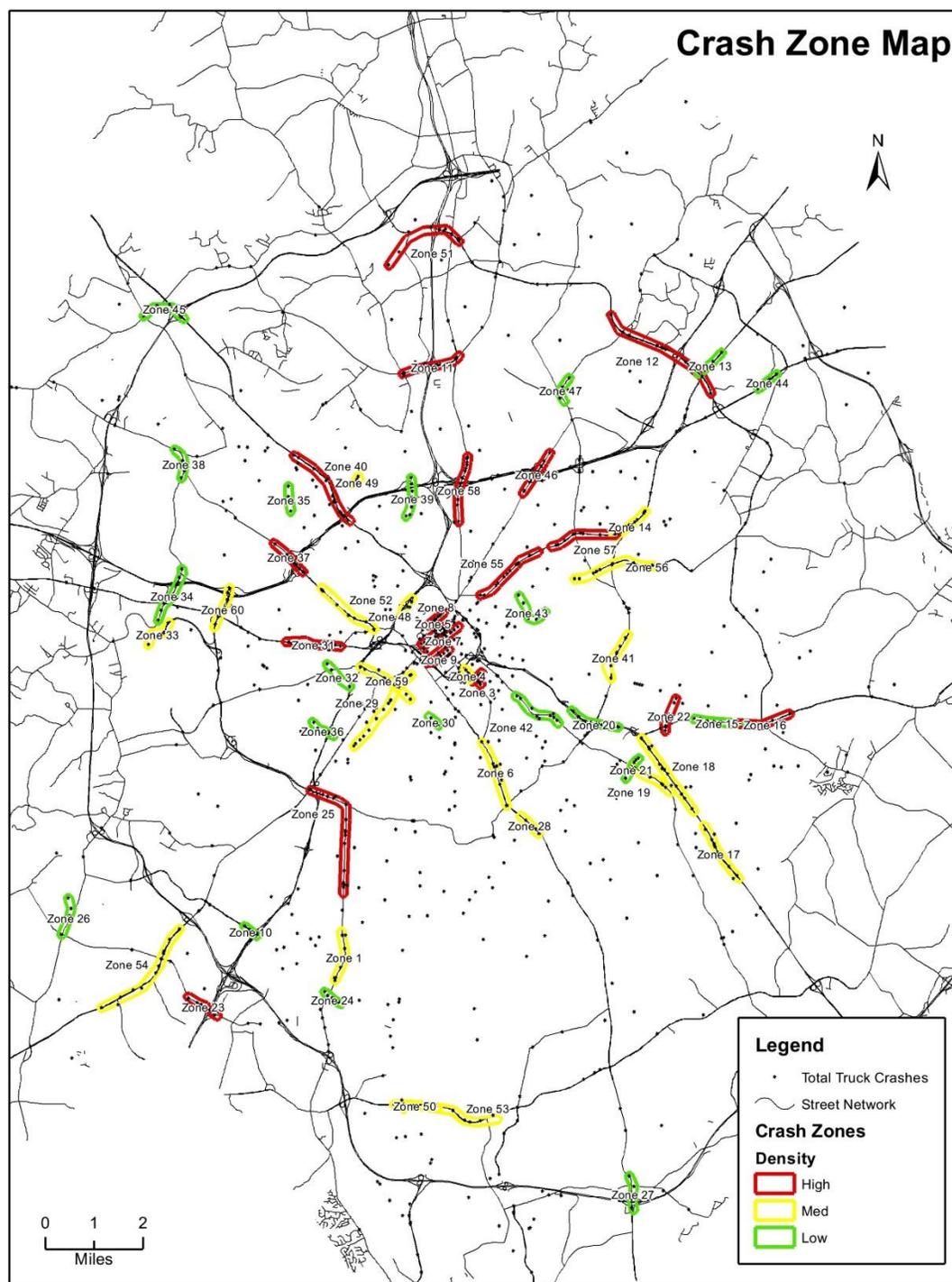
FIGURE12: Crash Zone Map

The zones identified from the Kernel density maps were ranked based on crash frequency, density and crash severity. A zone with say, 25 crashes, could rank 10 based on crash frequency while a zone with 4 crashes can rank 2based on the crash severity. Considering frequency, density and severity in the ranking may lead to unbiased results.

Table1 summarizes zone number (Zone), crash frequency (CF), rank based on crash frequency (Rank based on CF), crash density (CD), rank based on crash density (Rank based on CD), number of fatal, injury type A, B, C and PDO crashes, rank based on crash severity (Rank based on CS), average of sum of the ranks (Avg of sum of Ranks) and rank based on sum of the rank method (Rank).

TABLE 1: High Crash Zones Ranking

Zone	CF	Rank based on CF	CD	Rank based on CD	Fatal	Injury			PDO	CS score	Rank based on CS	Avg. of sum of ranks	Rank
						A	B	C					
1	4	52	22.71	55	0	0	0	1	3	4	53.00	53.33	56.00
2	14	15	274.33	2	0	0	0	2	12	14	26.00	14.33	10.00
3	7	37	124.35	11	0	0	1	0	5	24	15.00	21.00	22.00
4	27	1	453.77	1	0	0	2	2	23	63	1.00	1.00	1.00
5	19	5	176.38	5	0	0	0	5	0	5	44.00	18.00	16.00
6	11	21	48.16	36	0	0	0	11	0	11	31.00	29.33	31.00
7	3	58	21.15	56	0	0	0	2	1	3	59.00	57.67	59.00
8	4	48	50.1	34	0	0	0	0	4	4	50.00	44.00	47.00
9	19	5	193.71	4	0	0	1	0	18	37	5.00	4.67	3.00
10	5	44	75.42	24	0	0	0	0	5	5	45.00	37.67	41.00
11	6	41	28.96	49	0	0	0	0	6	6	43.00	44.33	48.00
12	16	10	37.42	42	0	0	1	4	11	34	6.00	19.33	20.00
13	6	41	52.18	33	0	0	0	1	5	6	41.00	38.33	42.00
14	4	48	27.83	51	0	0	0	0	4	4	54.00	51.00	54.00
15	7	37	43.72	39	0	0	0	2	4	6	42.00	39.33	43.00
16	6	41	33.39	45	0	0	1	0	5	24	16.00	34.00	35.00
17	19	5	88.4	18	0	0	0	4	15	19	21.00	14.67	12.00
18	4	48	13.62	58	0	0	0	3	1	4	56.00	54.00	58.00
19	15	11	91.85	17	0	0	0	7	8	15	23.00	17.00	15.00
20	8	30	45.33	38	0	0	0	1	7	8	37.00	35.00	37.00
21	3	55	32.08	46	0	0	1	1	1	21	18.00	39.67	44.00
22	7	37	55.35	29	0	0	0	1	6	7	40.00	35.33	39.00
23	5	44	39.06	41	0	0	0	0	5	5	46.00	43.67	46.00
24	8	30	97.82	15	0	0	1	0	7	26	13.00	19.33	19.00
25	11	21	26.81	52	0	0	0	0	11	11	32.00	35.00	38.00
26	20	4	145.06	7	0	0	0	2	18	20	19.00	10.00	6.00
27	11	21	87.25	19	0	0	0	1	10	11	30.00	23.33	26.00
28	8	30	83.44	22	0	0	0	1	7	8	36.00	29.33	30.00
29	15	11	49.04	35	0	0	0	1	14	15	25.00	23.67	27.00
30	8	30	136.07	8	0	0	1	1	6	26	12.00	16.67	14.00

TABLE 1: Continued

Zone	CF	Rank based on CF	CD	Rank based on CD	Crash severity (CS)					CS score	Rank based on CS	Avg. of sum of ranks	Rank
					Fatal	A	B	C	PDO				
31	15	11	78.54	23	0	0	1	0	14	33	7.00	13.67	9.00
32	8	30	67.69	26	0	0	0	2	2	4	49.00	35.00	36.00
33	12	18	93.94	16	0	0	0	1	11	12	29.00	21.00	23.00
34	11	21	55.29	30	0	0	1	1	9	29	10.00	20.33	21.00
35	12	18	125.84	10	0	0	0	3	9	12	28.00	18.67	17.00
36	3	55	30.83	47	0	0	0	0	3	3	58.00	53.33	55.00
37	5	44	34.33	43	0	0	0	0	5	5	47.00	44.67	49.00
38	4	48	33.65	44	0	0	0	1	3	4	51.00	47.67	51.00
39	4	48	28.45	50	0	0	0	0	4	4	52.00	50.00	53.00
40	4	48	13.64	58	0	0	0	0	4	4	55.00	53.67	57.00
41	19	5	107.25	12	0	0	0	0	19	19	20.00	12.33	7.00
42	13	16	70.21	25	0	0	0	1	12	13	27.00	22.67	25.00
43	16	10	104.5	14	0	0	2	4	10	52	2.00	8.67	5.00
44	22	3	246	3	0	0	0	4	18	22	17.00	7.67	4.00
45	9	28	86.56	20	0	0	0	1	8	9	35.00	27.67	29.00
46	18	9	106.97	13	0	0	0	5	13	18	22.00	14.67	11.00
47	9	28	85.66	21	0	0	0	1	8	9	34.00	27.67	28.00
48	11	21	135.77	9	0	0	1	1	9	29	9.00	13.00	8.00
49	3	55	58.49	28	0	0	0	1	2	3	57.00	46.67	50.00
50	8	30	26.66	53	0	0	0	1	7	8	39.00	40.67	45.00
51	7	37	30.7	48	0	0	1	0	6	25	14.00	33.00	34.00
52	10	26	46.07	37	0	0	0	3	7	10	33.00	32.00	33.00
53	10	26	26.6	54	0	0	1	3	6	28	11.00	30.33	32.00
54	5	44	20.17	57	0	0	0	1	4	5	48.00	49.67	52.00
55	15	11	52.93	32	0	0	0	1	14	15	24.00	22.33	24.00
56	14	15	61.47	27	0	1	1	2	10	50	3.00	15.00	13.00
57	12	18	53.1	31	0	0	1	2	9	30	8.00	19.00	18.00
58	8	30	39.07	40	0	0	0	0	8	8	38.00	36.00	40.00
59	25	2	168.3	6	0	0	1	3	21	43	4.00	4.00	2.00

From the table it can observed that ranks vary based on the individual method considered.

For example, Zone 2is ranked15 based on crash frequency, 2 based on crash density but

26 based on crash severity. Therefore, to minimize the demerits of individual methods,

sum of the ranks method was used to compute combined rank based on the three individual methods. The sum of the rank was attained by summing the ranks based on frequency, density and severity and then computing the average. The zones were thus ranked based on sum of ranks method.

4.4 On- and Off-network Characteristics

Buffers of 0.25-, 0.5- and 1-mile width are generated around each of the 60 linear zones. Due to inconsistency in the zone characteristics and lack of required data, zones 25 and 51 (though shown in Figures 13 - 27) were not included for further analysis.

Land use, socio-economic and demographic data are overlaid over these buffers. The network characteristics as speed limit, the number of lanes, the number of driveways, horizontal curvature of the zones, presence of median were gathered from Google maps and Google earth. In addition, traffic volume (based on outputs from calibrated four-step regional travel demand forecasting models) and segment length were also obtained and added to the database.

Maps generated with various buffers around each linear zone are shown in figures 13 to 15.

Once the buffers of 0.25-, 0.5- and 1-mile were generated, the land use, demographic and socio-economic data layers were overlaid onto the buffers. Figure 16 to 27 are some maps with land use, socio–economic and demographic data overlaid on to the buffers around the high truck locations and land use data.

.25 Mile Buffer from Linear Zone
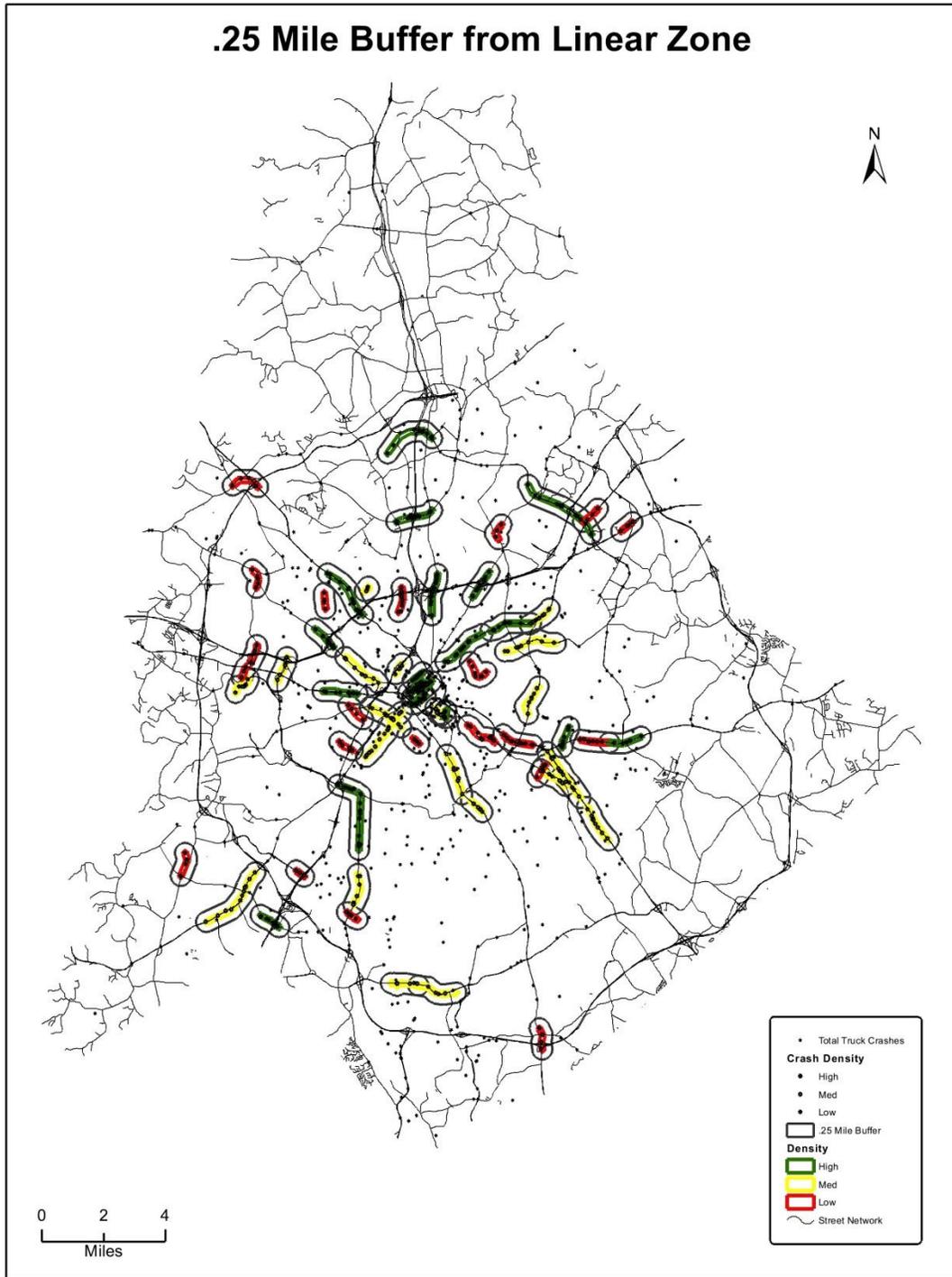
FIGURE 13: 0.25-mile Radius from Linear Zones
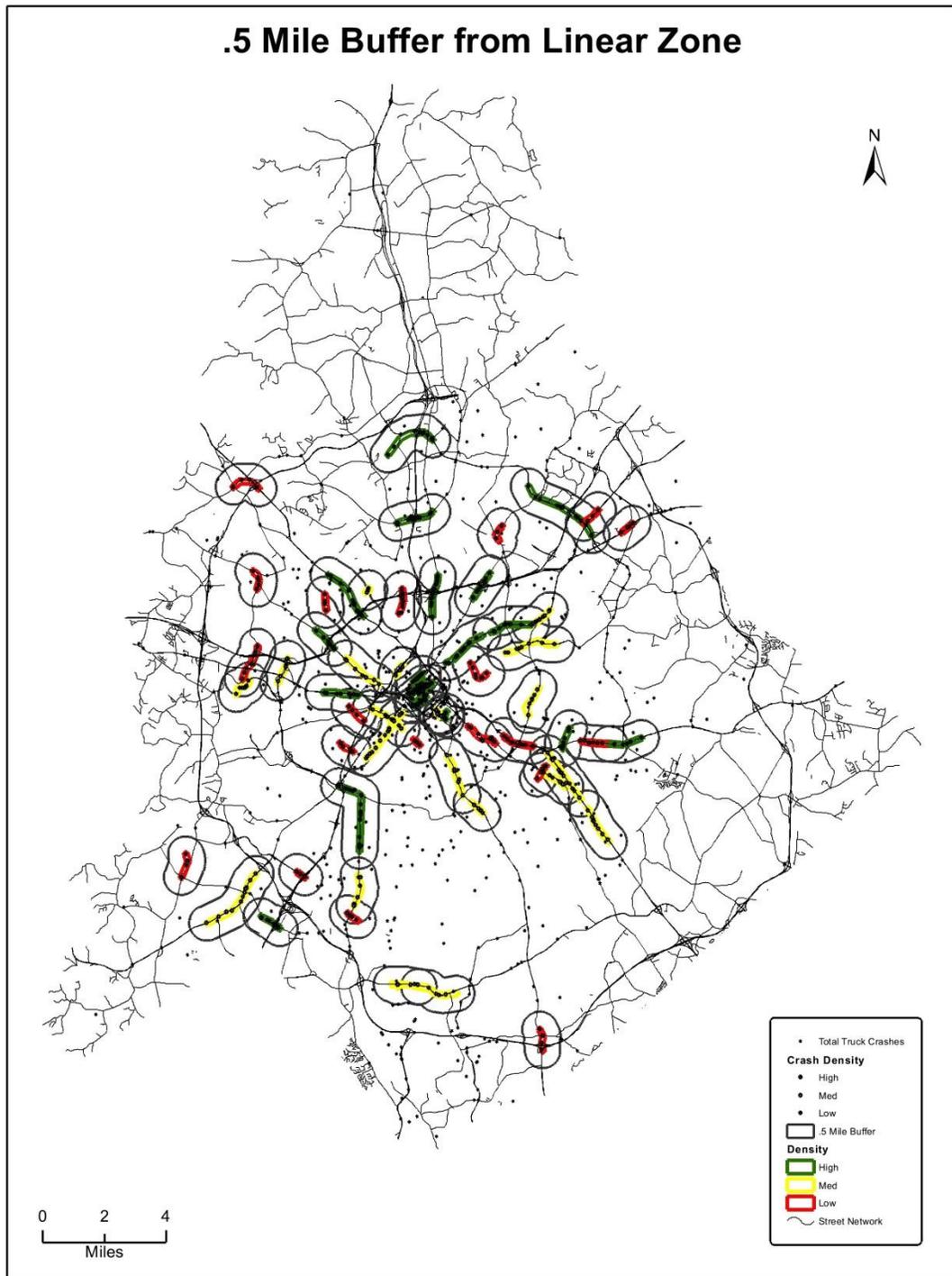
FIGURE 14: 0.5-mile Radius from Linear Zones
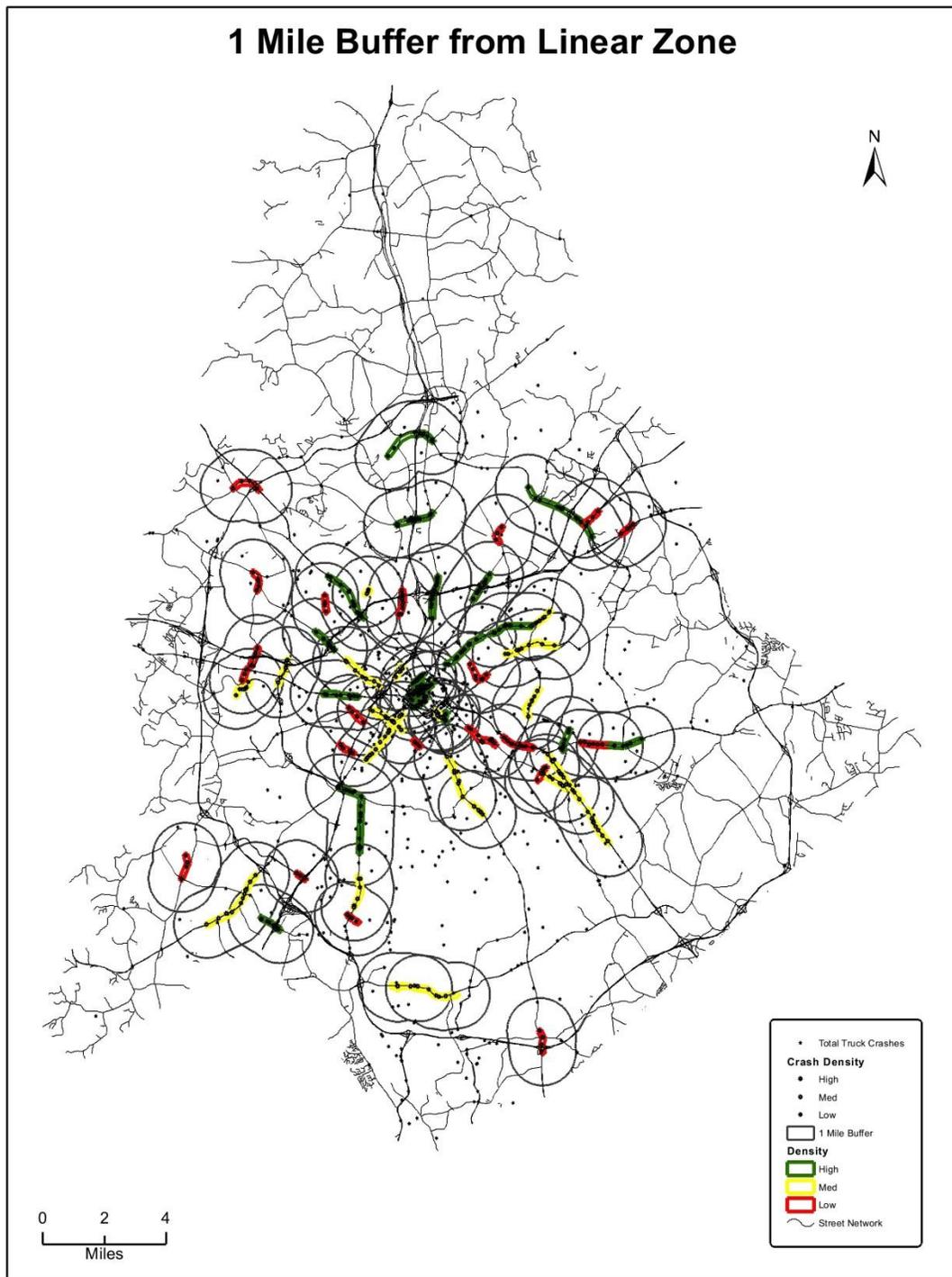
FIGURE15: 1-mile Radius from Linear Zones
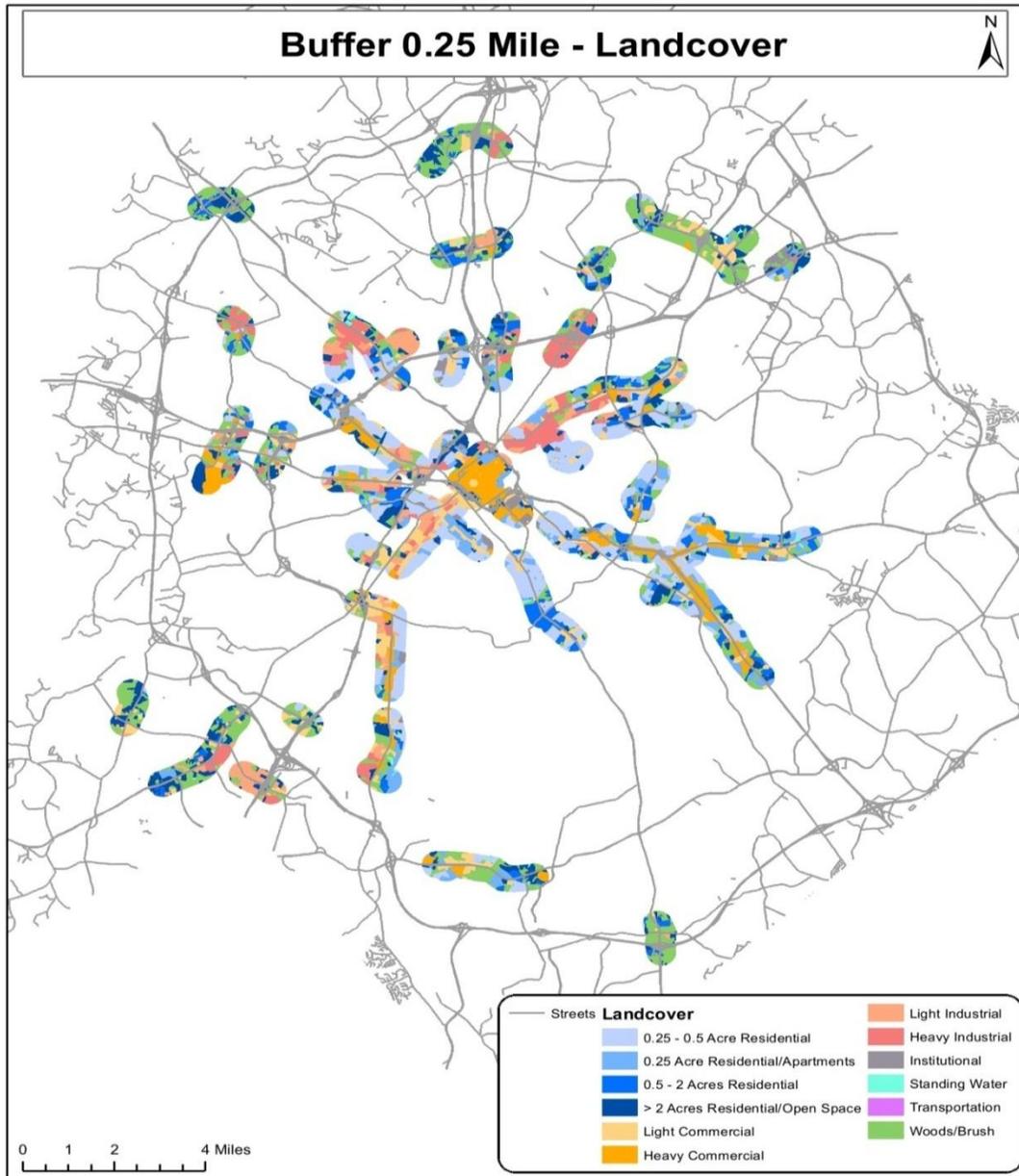
FIGURE 16: Land Use Data Overlaid onto 0.25-mile Buffers
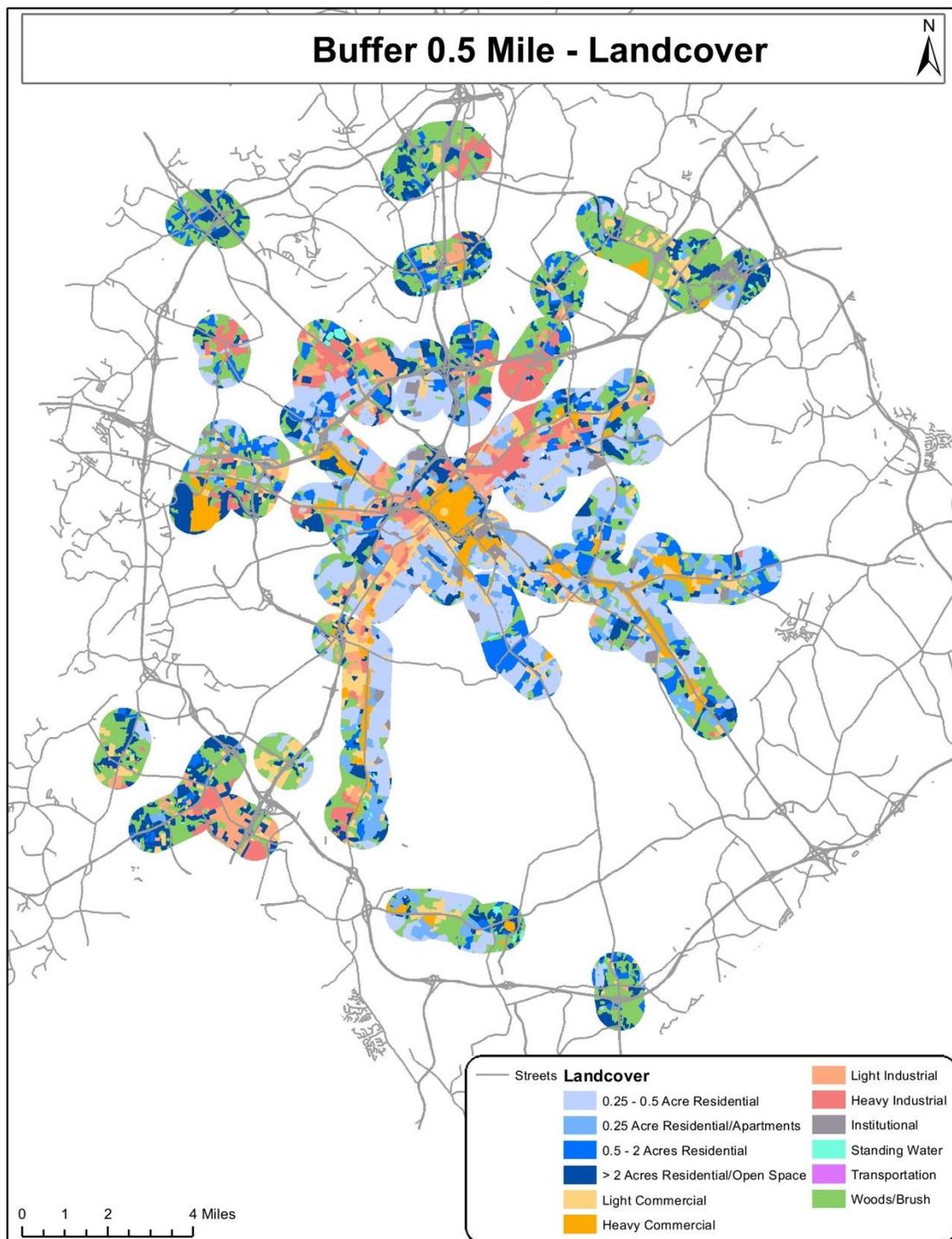
FIGURE 17: Land Use Data Overlaid onto 0.5-mile Buffers
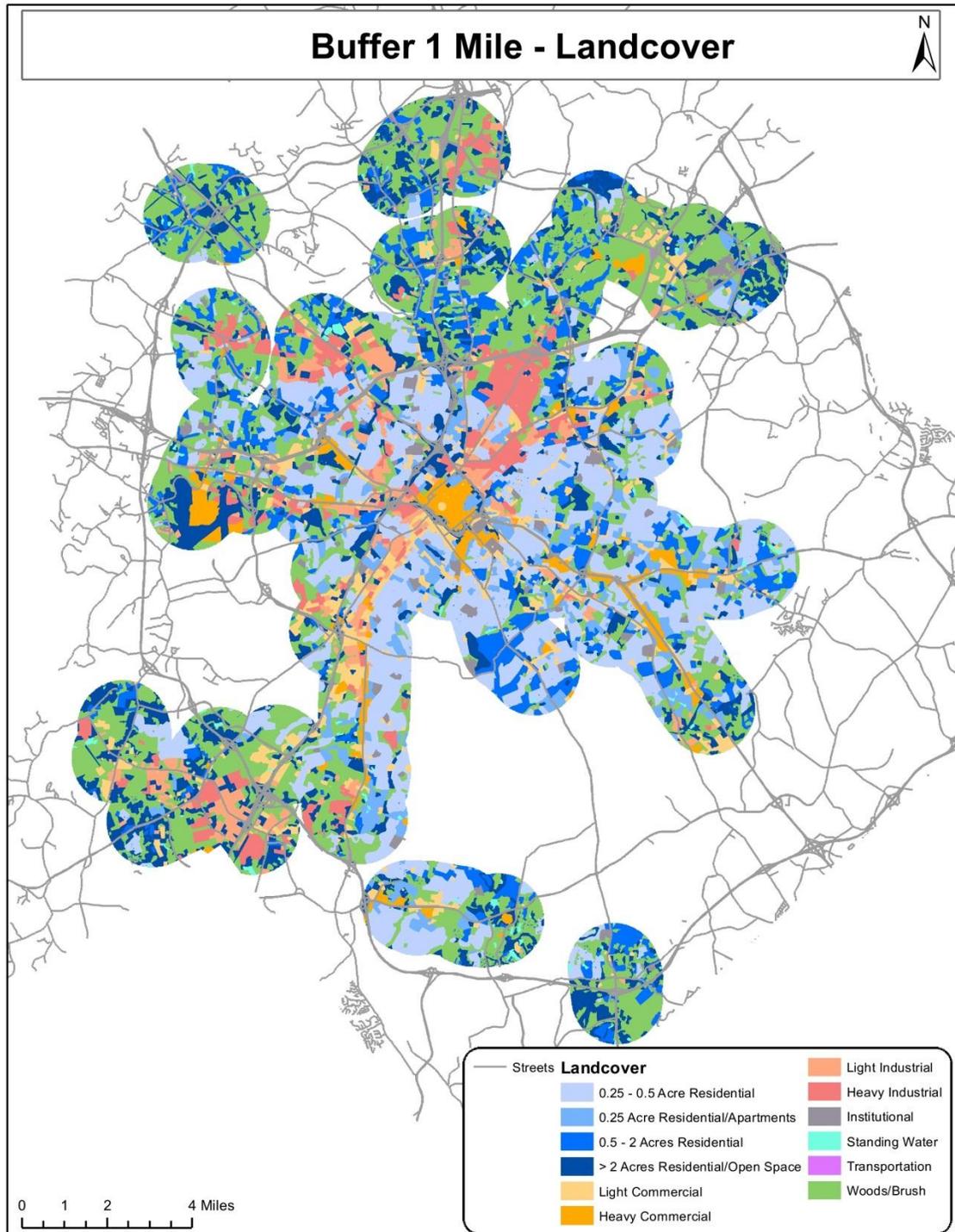
FIGURE 18: Land Use Data Overlaid onto 1.0-mile Buffers
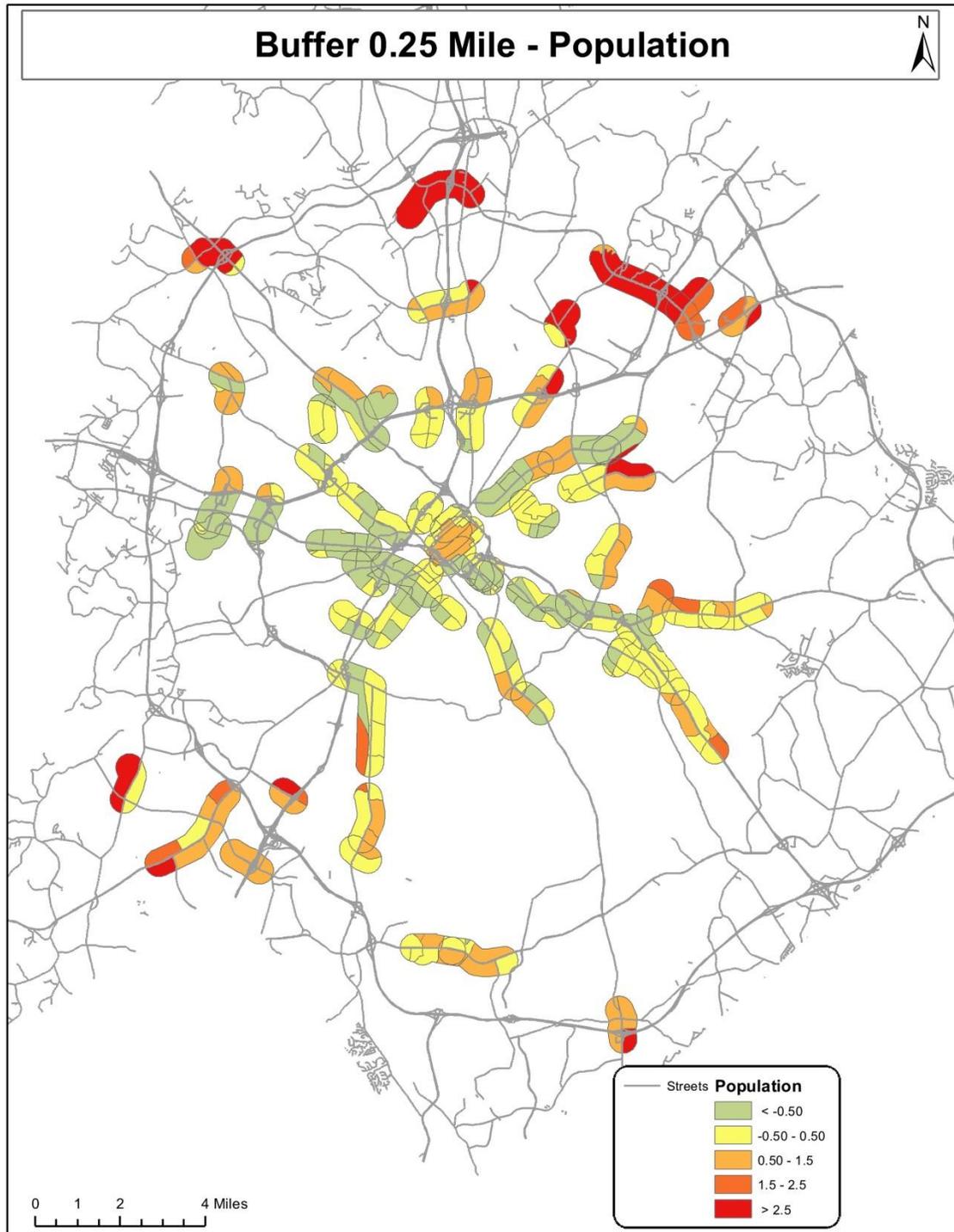
FIGURE19: Population Data Overlaid onto 0.25-mile Buffers

Buffer 0.25 Mile - Unemployment Rate

N

Streets	**Unemployment Rate**
	< -0.50
	-0.50 - 0.50
	0.50 - 1.5
	1.5 - 2.5
	> 2.5

0   1   2   4 Miles
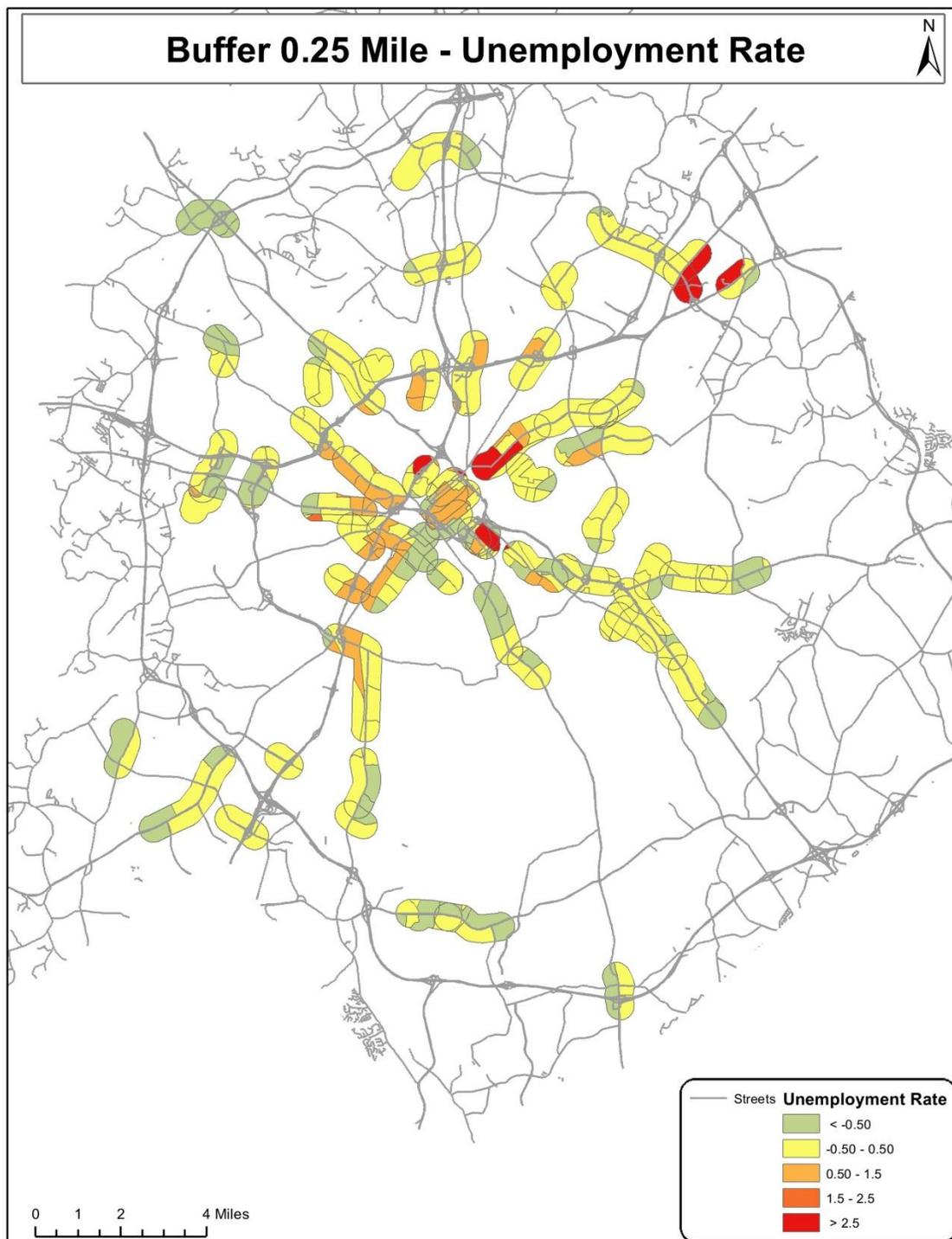
FIGURE20: Unemployment Data Overlaid onto 0.25-mile Buffers
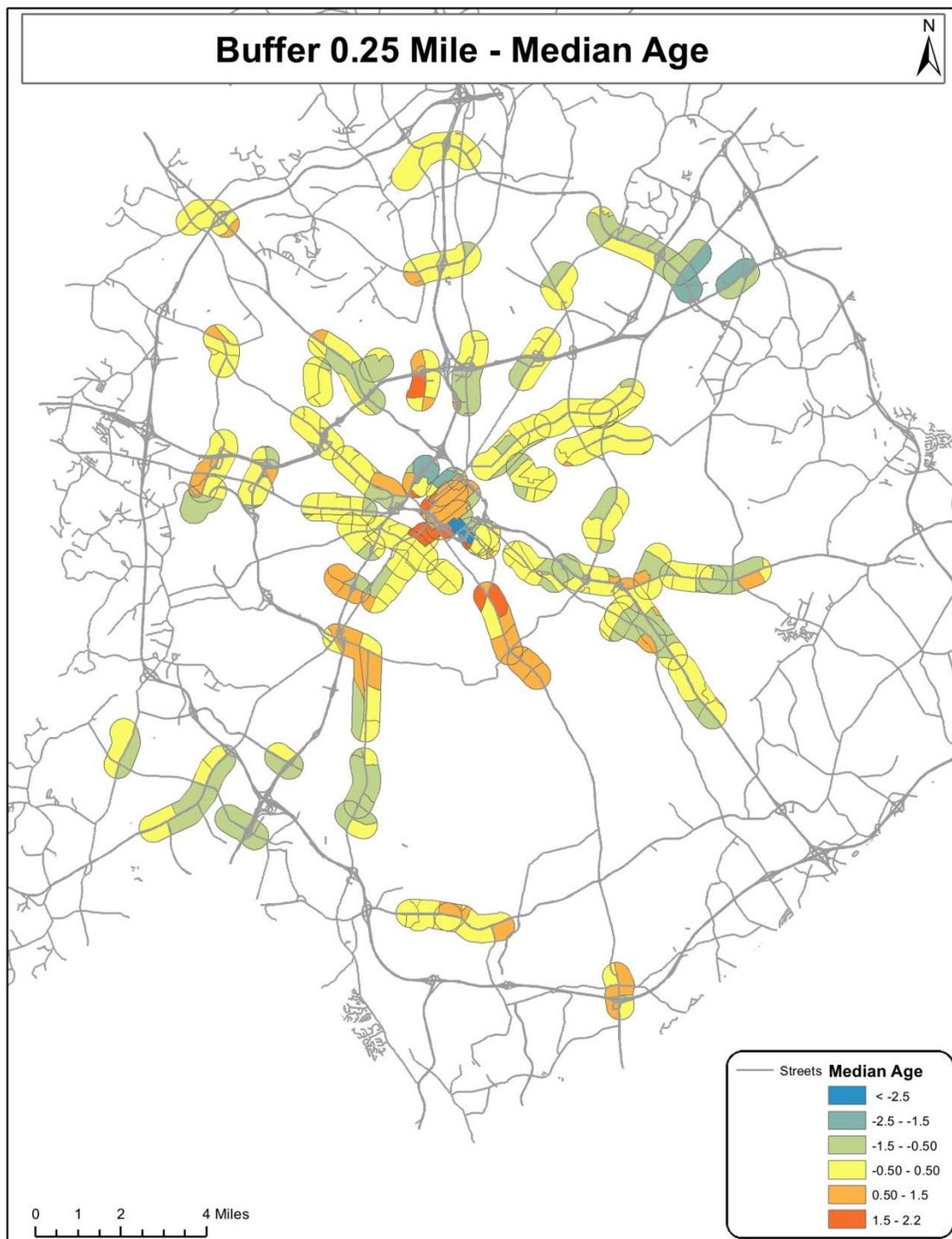
FIGURE21: Median Age Data Overlaid onto 0.25-mile Buffers
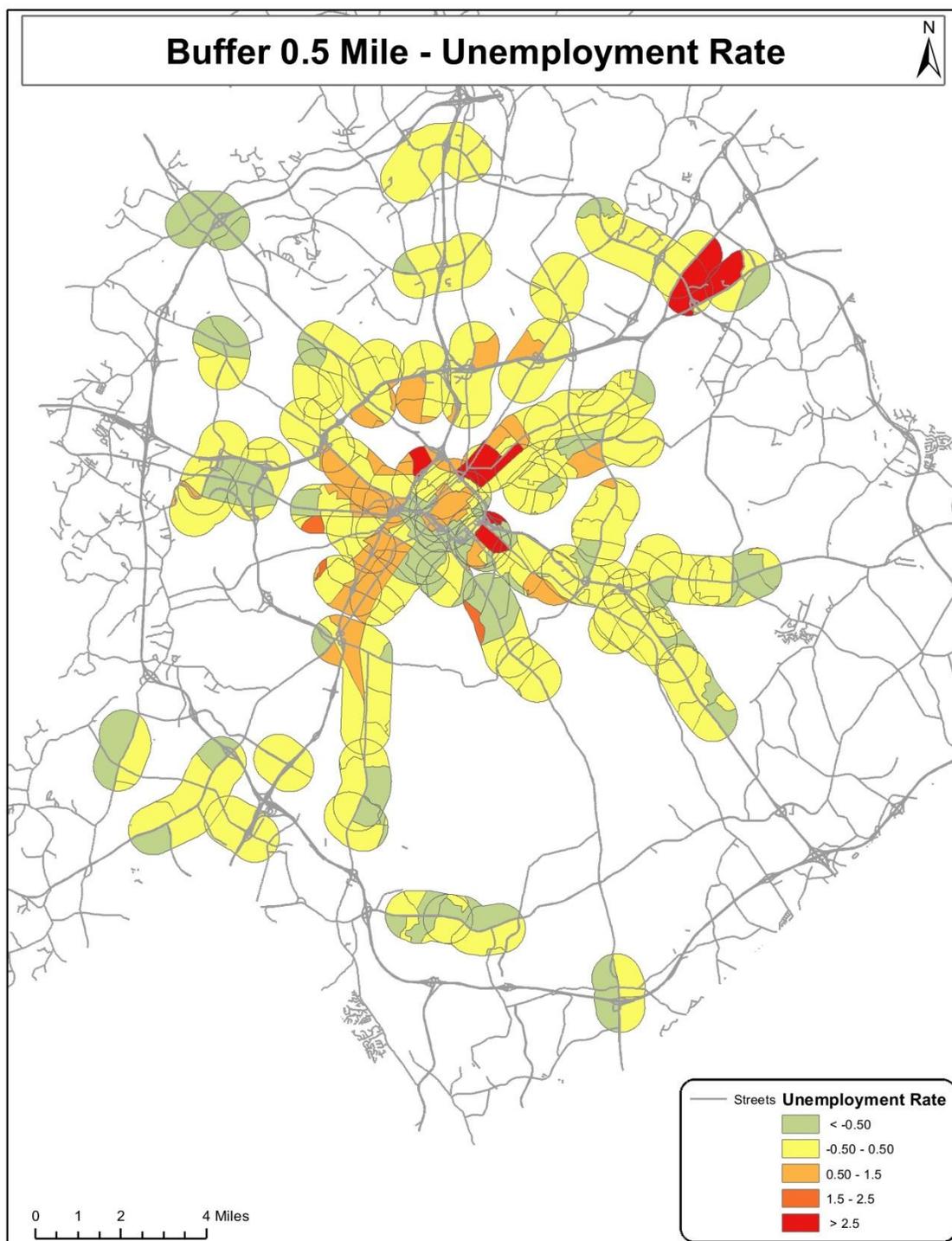
FIGURE22: Unemployment Data Overlaid onto 0.5-mile Buffers
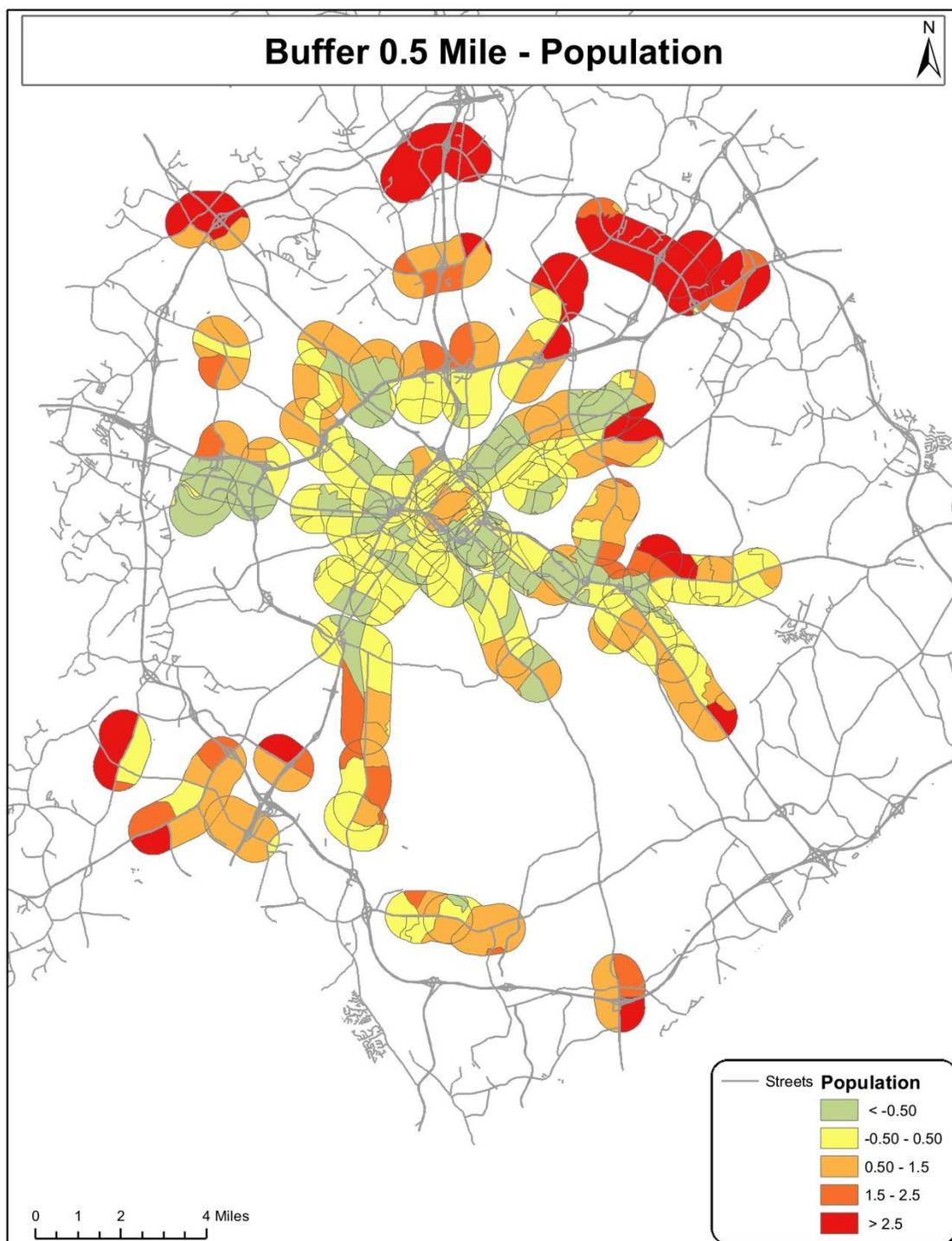
FIGURE23: Population Data Overlaid onto 0.5-mile Buffers
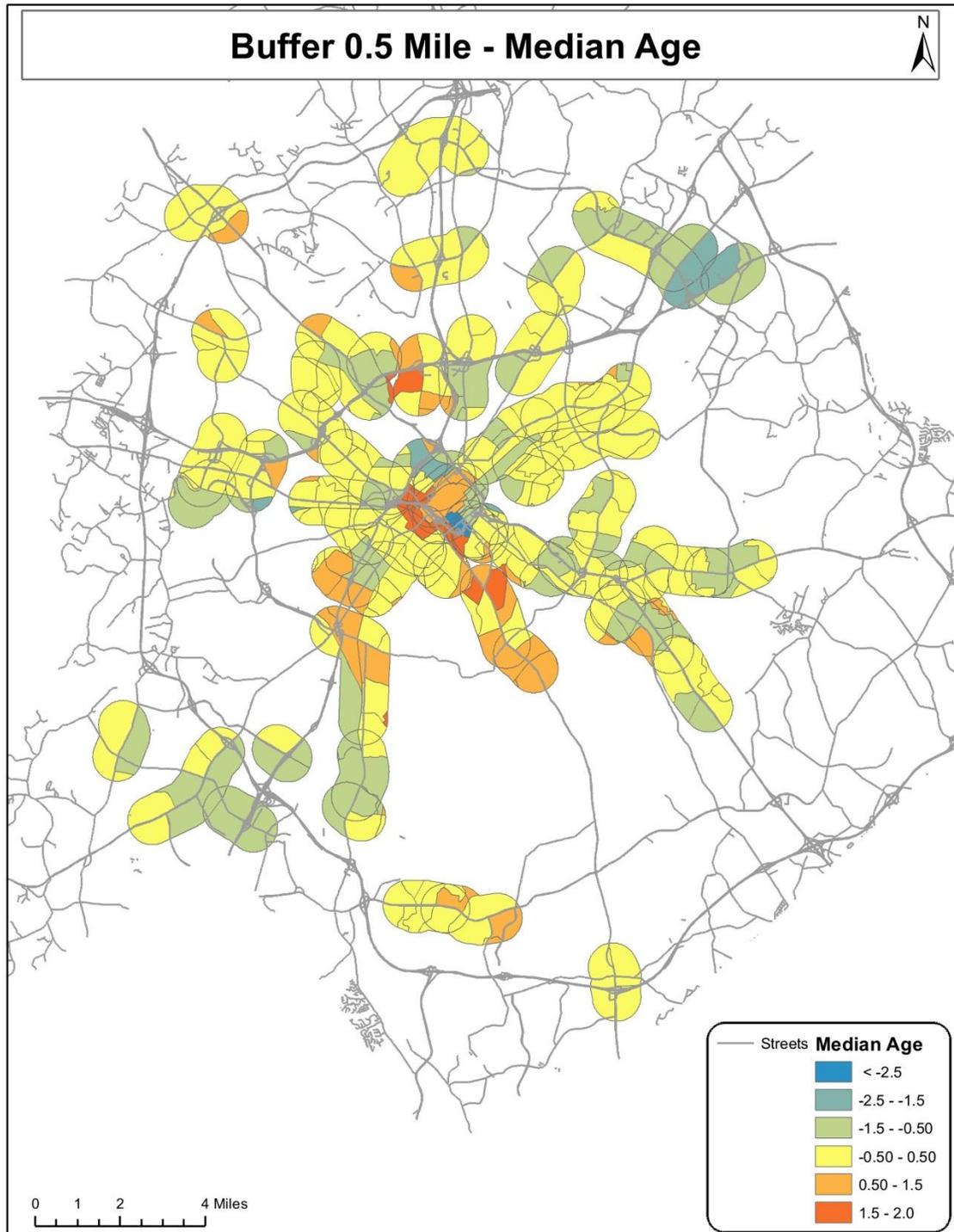
FIGURE24: Median Age Data Overlaid onto 0.5-mile Buffers
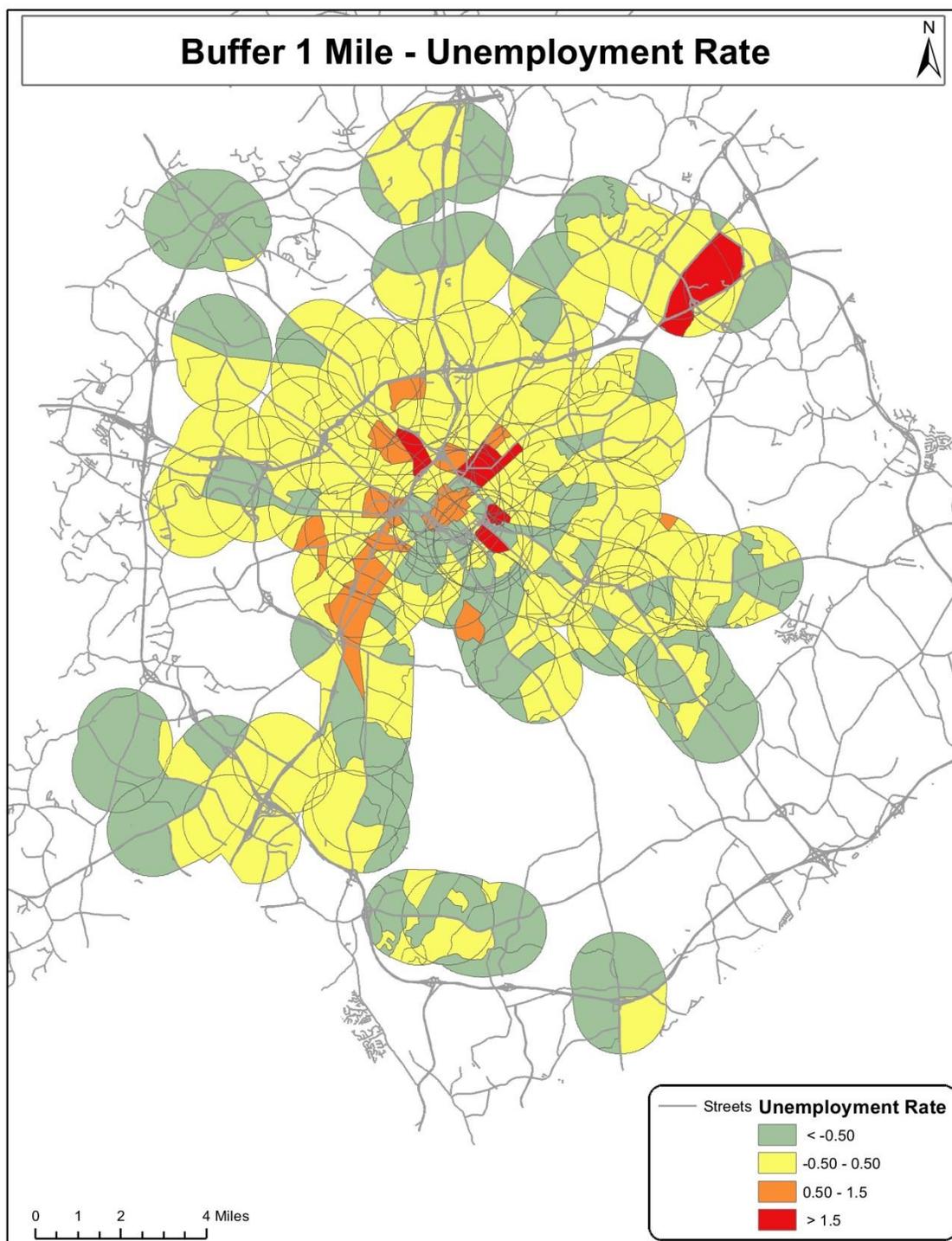
FIGURE25: Unemployment Data Overlaid onto 1.0-mile Buffers
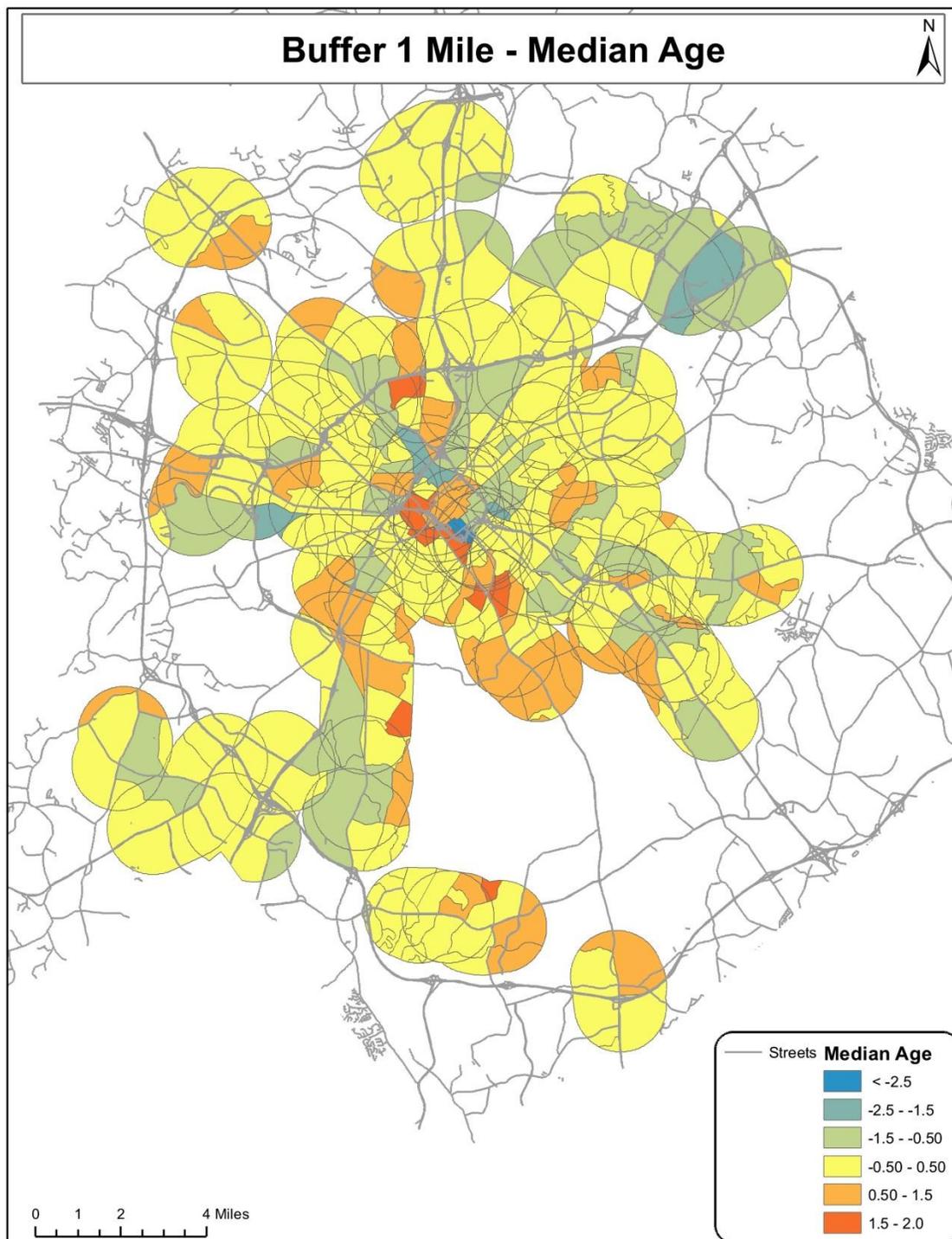
FIGURE26: Median Age Data Overlaid onto1.0-mile Buffers
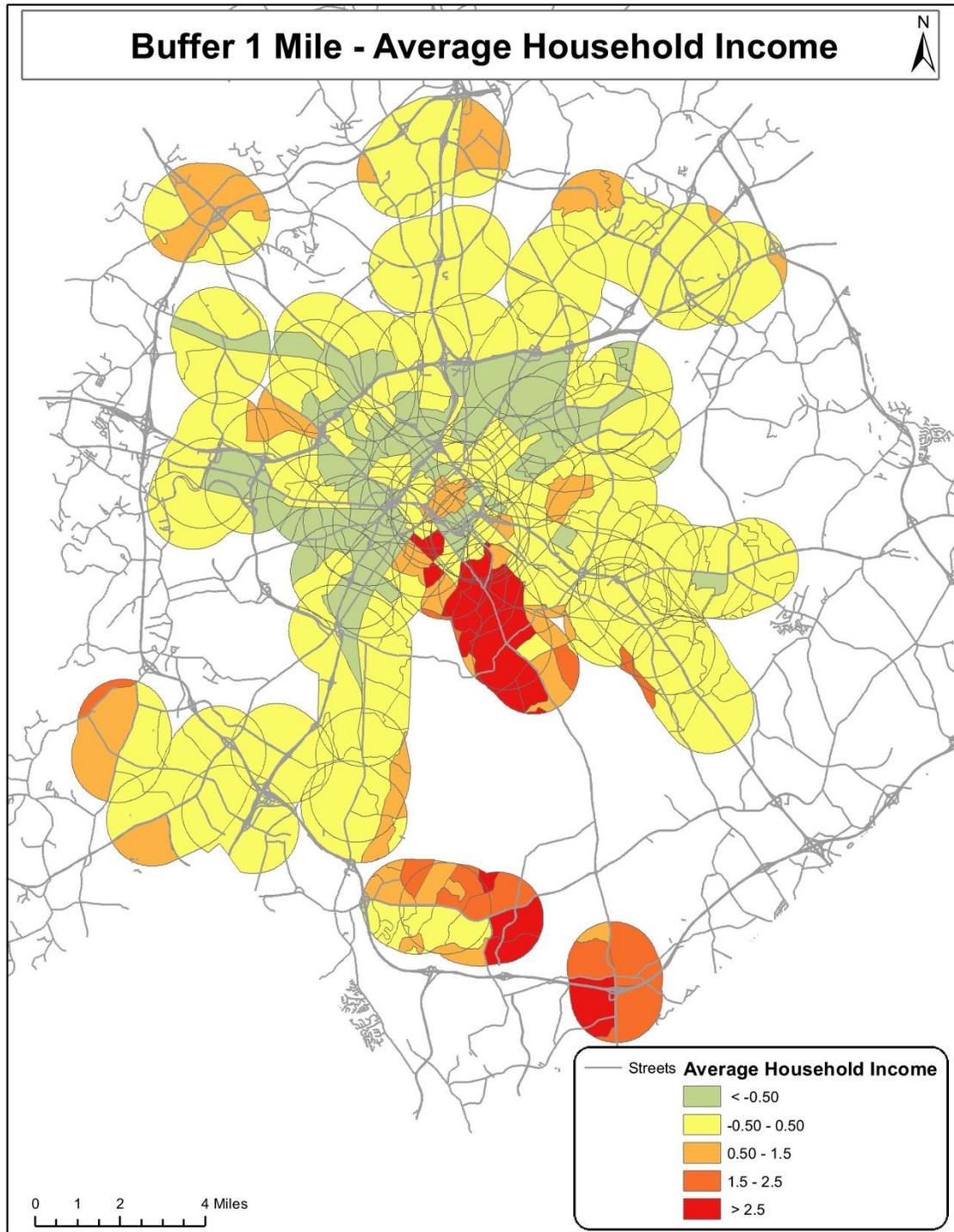
FIGURE27: Average Household Data Overlaid onto 1.0-mile Buffers

After overlapping various network data onto the buffers, the variables from each buffer were extractedTable2lists variables considered for analysis in this research.

TABLE2: Independent Variables Used in this Research

S. No.	Code	Explanation
1	R1	0.25 - 0.5 Acre Residential
2	R2	0.25 Acre Residential/Apartments
3	R3	0.5 - 2 Acres Residential
4	R4	2 Acres Residential/Open Space
5	HC	Heavy Commercial
6	HI	Heavy Industrial
7	I	Institution
8	LC	Light Commercial
9	LI	Light Industrial
10	POP	Population
11	POPFAM	Population in Family Households
12	AGEMED	Median Age
13	AGEMEDM	Median Age Males
14	AGEMEDF	Median Age Females
15	VPH	Auto Ownership
16	AHHI	Average Household Income
17	UNEMR	Unemployment Rate
18	LBF	Population in Labor Force
19	Lanes	Number of lanes
20	Driveways	Number of driveways
21	Signals	Number of Signalized Intersections
22	HorizCurv	Number of Horizontal curvatures
23	Median	Median
24	Zone Length	Zone Length
25	AADT	Average Annual Daily Traffic

After summarizing the variables extracted from each buffer width, it was found that residential, commercial and industrial spaces were high in the 1-mile buffer width zones while least in the 0.25-mile buffer width zones. Similarly, population and average

household income was observed to be high in the 1-mile buffer width when compared to 0.25-mile buffer zones. Intuitively from the figures, it can be said variables in 0.5- and 1-mile buffer width zones might have more influence compared to 0.25-mile buffer width variables.

4.5 Crash Estimation Models

As discussed in Chapter 3, demographic, socio-economic, network, and land use characteristics for 50 randomly selected zones (of the 60 identified using Kernel density maps) were used to develop crash estimation models.

Crash estimation models were developed using the final variables identified from the correlation matrices for multiple cases. For each buffer width and different crash density categories, different crash estimation models were developed. The final recommended model was selected based on the goodness of statistics.

4.6 Correlation Matrices and Final Variables

To observe the effect of spatial proximity in capturing data and developing models, this research used 0.25-, 0.5- and 1-mile buffer widths to extract off-network characteristics data. Correlation matrices were developed to evaluate the strength of correlation between the independent variables for data based on all the three buffer widths.

As discussed in Chapter 3, the strength of the linear association between two variables was determined by the correlation coefficient between the two variables. A correlation matrix was generated using the SPSS® software. Independent variables were screened such that no final selected variables have correlations out of (-0.3, +0.3) range. Variable selection was carried out in a logical way. Among the four types of independent

variables (network, demographic, socio-economic, and land use variables), the ones that are expected to be directly related to crash count were included whereas the variables that are comparatively unrelated were excluded. For example, in the network characteristics data, the more number of lanes, the more traffic and exposure and hence, the probability of more collisions. Speed limit also has a great influence on traffic and crash counts.Therefore, speed limit was considered along with the number of lanes wherever applicable. In the land use characteristics data, elimination was carried out on a case-by-case basis. The objective was to eliminate the effect of multicollinearity to the extent possible and to avoid over parameter zing the final model.

The above process is repeated for all selected buffer widths datasets considered in this research. The final sets of independent variables (that are not correlated to each other) are selected to develop the models for each buffer width.

4.6.1.Correlation Matrix and Final List of Variables for 0.25-mile Buffer Width Dataset

The correlation matrix for 0.25-mile buffer width data was developed to observe the correlation between the independent variables (land use, demographics,socio-economic and network characteristics, etc.).Final variables identified from the correlation matrix for this buffer width dataset are shown in Table3.

The red colored numbers in the matrix (Table3) denote the variables with correlation coefficients that are unacceptable. Table4 lists final set of variables for this buffer width dataset.

TABLE3: Correlation Matrix - 0.25-mile Buffer

Variables	R1	R2	R3	R4	HC	HI	I	LC	LI	POP	POP FAM	AGE MED	AGE MED M	AGE MED F	VPH	AHHI	LBF	UNE MR	Lanes	Drive ways	Signals	Median	Horiz Curv	Zone Length	AADT
R1	1																								
R2	-0.06	1																							
R3	0.17	-0.1	1																						
R4	-0.18	-0.3	0.16	1																					
HC	-0.04	-0	-0.22	-0.22	1																				
HI	-0.06	0.14	0.17	-0.22	-0.35	1																			
I	-0.2	-0.3	-0.06	-0.04	-0.13	-0.45	1																		
LC	0.2	0.09	-0.14	0.06	-0.09	-0.1	-0.23	1																	
LI	0.31	-0.1	-0.35	-0.13	-0.18	-0.08	0.1	0.27	1																
POP	0.65	0.47	0.06	-0.12	0.09	-0.09	-0.29	0.44	-0	1															
POPFAM	0.76	0.41	0.17	-0.04	-0.12	-0.05	-0.31	0.36	0.03	0.92	1														
AGEMED	0.76	0.12	0.05	-0.38	0.18	-0.04	-0.08	0.24	0.25	0.65	0.6	1													
AGEMEDM	0.74	0.12	0.06	-0.38	0.17	-0.04	-0.05	0.24	0.25	0.63	0.59	1	1												
AGEMEDF	0.77	0.12	0.05	-0.38	0.2	-0.05	-0.12	0.24	0.24	0.66	0.61	1	0.99	1											
VPH	0.61	0.46	0.14	-0.04	-0.01	-0.13	-0.34	0.49	-0.1	0.91	0.86	0.53	0.52	0.54	1										
AHHI	0.6	0.09	0.31	-0.38	0.06	-0.04	-0.01	0.16	0.18	0.45	0.43	0.76	0.77	0.746	0.54	1									
LBF	0.3	-0.1	-0.17	-0.3	0.21	0.158	0.32	0.08	0.29	0.29	0.22	0.73	0.74	0.713	0.05	0.35	1								
UNEMR	0.56	0.5	0	-0.12	0.13	-0.12	-0.29	0.47	-0.1	0.98	0.85	0.59	0.57	0.602	0.92	0.41	0.248	1							
Lanes	0.2	0.2	0.1	-0.07	0.1	-0.07	-0.24	0.13	-0	0.35	0.34	0.31	0.31	0.313	0.23	0.132	0.205	0.33	1						
Driveways	0.56	-0	0.5	-0.15	-0.3	-0.19	-0.04	-0.1	-0.1	0.27	0.33	0.43	0.44	0.422	0.34	0.755	0.046	0.2	-0.09	1					
Signals	0.14	0.08	-0.05	0.162	0.33	0.21	-0.22	0.55	0.17	0.47	0.3	0.26	0.25	0.279	0.45	0.187	0.207	0.5	-0.02	-0.05	1				
Median	0.36	0.17	0.16	0.228	-0.34	0.073	-0.18	0.3	0.17	0.25	0.39	0.05	0.04	0.05	0.26	0.03	-0.13	0.2	0.276	0.133	0.112	1			
Horiz Curv	0.03	-0.2	0.15	0.622	-0.39	0.19	-0.28	0.37	0.01	0.16	0.2	-0.12	-0.11	-0.12	0.25	-0.059	-0.15	0.18	-0.01	-0.02	0.334	0.2439	1		
Zone Length	0.45	0.03	0.18	0.432	-0.1	0.243	-0.39	0.61	0.23	0.54	0.61	0.24	0.23	0.238	0.59	0.165	0.034	0.51	0.143	0.059	0.592	0.4555	0.708	1	
AADT	0.05	-0.2	0.05	0.32	0.31	0.229	-0.31	0.29	0.22	0.19	0.14	-0.02	-0.02	-0	0.16	-0.088	-0.04	0.2	0.256	-0.16	0.486	0.0925	0.531	0.6081	1

Note: R1= 0.25 - 0.5 Acre Residential, R2= 0.25 Acre Residential/Apartments, R3=0.5 - 2 Acres Residential, R4=> 2 Acres Residential/Open Space, HC=Heavy Commercial, HI=Heavy Industrial, I=Institutional,LC=Light Commercial, LI=Light Industrial, POP=Population, POPFAM=Population in Family Households, AGEMED=Median Age, AGEMEDM=Median Age Males, AGEMEDF=Median Age Females,VPH represents ownership, AHHI=Average household Income, UNEMR=Unemployment Rate, LBF Population In Labor Force, Lanes = Number of lanes, Driveways = Number of driveways, Signals= Number of signalized intersections, HorizCurv= Number of horizontal curvatures , Median= Median

TABLE4: Final Variables - 0.25-mile Buffer

Variables	Description
R1	0.25 - 0.5 Acre Residential
R2	0.25 Acre Residential/Apartments
R3	0.5 - 2 Acres Residential
LC	Light Commercial
HC	Heavy Commercial
Lanes	Number of lanes
I	Industrial

4.6.2. Correlation Matrix and Final List of Variables for 0.5-mile Buffer Width Dataset

As explained for the 0.25-mile buffer scenario, correlation matrix for this for 0.5-mile buffer width data was developed to observe the correlation between the independent

variables (land use, demographics, socio-economic and network characteristics).Table5shows the correlation matrix and Table6shows the final set of variables identified from the correlation matrix for 0.5-mile buffer width dataset. Variables with correlation values within the acceptable range are selected for further analysis and rest are excluded.

TABLE5: Correlation Matrix - 0.5-mile Buffer

Variables	R1	R2	R3	R4	HC	HI	I	LC	LI	POP	POP FAM	AGE MED	AGE MED M	AGE ME DF	VPH	AHHI	LBF	UN EM R	Lanes	Drive ways	Signals	Median	Horiz Curv	Zone Length	AADT
R1	1																								
R2	0.1	1																							
R3	0.29	-0.11	1																						
R4	-0.4	-0.39	-0	1																					
HC	-0.2	-0.04	-0.4	-0.01	1																				
HI	-0	-0.21	-0.1	-0.1	-0.23	1																			
I	0.08	0.12	-0.3	-0.02	0.141	-0.33	1																		
LC	0.25	0.25	-0.2	-0.05	-0	-0.11	-0.03	1																	
LI	0.24	-0.08	-0.2	0.05	-0.09	0.221	-0.22	0.19	1																
POP	0.61	0.48	0.11	-0.25	0.073	-0.15	0.11	0.41	-0	1															
POPFAM	0.73	0.4	0.27	-0.21	-0.19	-0.05	0.03	0.29	-0	0.92	1														
AGEMED	0.68	0.3	0.03	-0.42	0.103	-0.09	0.2	0.29	0.25	0.64	0.569	1													
AGEMEDM	0.67	0.31	0.03	-0.43	0.095	-0.09	0.22	0.29	0.24	0.63	0.558	1	1												
AGEMEDF	0.68	0.29	0.02	-0.42	0.131	-0.09	0.18	0.29	0.25	0.66	0.575	1	0.99	1											
VPH	0.56	0.49	0.25	-0.21	-0.08	-0.2	0.01	0.4	-0.1	0.9	0.866	0.48	0.48	0.48	1										
AHHI	0.54	0.22	0.31	-0.39	-0.13	0.13	0.19	0.2	0.45	0.393	0.77	0.79	0.76	0.51	1										
LBF	0.3	0.1	-0.2	-0.25	0.132	0.158	0.36	0.2	0.26	0.35	0.264	0.79	0.8	0.79	0.06	0.429	1								
UNEMR	0.52	0.52	0.05	-0.23	0.094	-0.19	0.12	0.45	-0.1	0.98	0.861	0.58	0.57	0.59	0.91	0.415	0.28	1							
Lanes	0.2	0.25	0.08	-0.07	-0.05	-0.17	0.07	0.13	-0	0.38	0.389	0.29	0.29	0.29	0.26	0.121	0.22	0.36	1						
Driveways	0.48	-0.03	0.54	-0.27	-0.32	-0.14	0.02	-0.1	-0	0.23	0.287	0.4	0.4	0.38	0.3	0.714	0.1	0.18	-0.09	1					
Signals	0.12	0.04	-0.1	0.2	0.252	0.122	-0.13	0.59	0.28	0.46	0.303	0.28	0.26	0.3	0.41	0.192	0.2	0.47	-0.02	-0.05	1				
Median	0.34	0.06	0.21	0.23	-0.52	0.153	-0.1	0.17	0.08	0.19	0.349	-0.03	-0.04	-0	0.23	-0.024	-0.1	0.16	0.276	0.133	0.1121	1			
Horiz Curv	-0.1	-0.15	0.19	0.55	-0.25	0.142	-0.2	0.26	-0	0.14	0.191	-0.15	-0.14	-0.2	0.25	-0.081	-0.1	0.17	-0.01	-0.02	0.3341	0.24389	1		
Zone Length	0.38	-0.02	0.29	0.36	-0.19	0.231	-0.2	0.5	0.22	0.51	0.592	0.17	0.17	0.17	0.57	0.121	0.03	0.49	0.143	0.059	0.5921	0.45555	0.708	1	
AADT	0.07	-0.22	0.1	0.35	0.254	0.111	-0.16	0.28	0.2	0.19	0.154	-0.04	-0.05	-0	0.16	-0.113	-0.1	0.19	0.256	-0.16	0.4855	0.0925	0.531	0.6081	1

Note: R1= 0.25 - 0.5 Acre Residential, R2= 0.25 Acre Residential/Apartments, R3=0.5 - 2 Acres Residential, R4=> 2 Acres Residential/Open Space, HC=Heavy Commercial, HI=Heavy Industrial, I=Institutional,LC=Light Commercial, LI=Light Industrial, POP=Population , POPFAM=Population in Family Households, AGEMED=Median Age, AGEMEDM=Median Age Males, AGEMEDF=Median Age Females,VPH represents ownership, AHHI=Average household Income, UNEMR=Unemployment Rate, LBF Population In Labor Force, Lanes = Number of lanes, Driveways = Number of driveways, Signals= Number of signalized intersections, HorizCurv= Number of horizontal curvatures , Median= Median

TABLE 6: Final variables - 0.5-mile buffer

Variables	Description
R1	0.25 - 0.5 Acre Residential
R2	0.25 Acre Residential/Apartments
R3	0.5 - 2 Acres Residential
LC	Light Commercial
LI	Light Industrial
Lanes	Number of lanes
HI	High Industrial
AADT	Average Annual Daily Traffic

4.6.3. Correlation Matrix and Final List of Variables for1-mile Buffer Width Dataset

In this case, the correlation matrix for 1.0-milebufferwidth data was developed to observe the correlation between the independent variables (land use, demographic, socio-economic and network characteristics). Table7shows the correlation matrix and

Table8shows the final set of variables identified from the correlation matrix for 1.0-mile

buffer width dataset.

TABLE7: Correlation Matrix - 1.0-mile Buffer

Variables	R1	R2	R3	R4	HC	HI	I	LC	LI	POP	POPFAM	AGEMED	AGEMEDM	AGEMEDF	VPH	AHHI	LBF	UNEMR	Lanes	Driveways	Signals	Median	HorizCurv	ZoneLength	AADT
R1	1																								
R2	0.5	1																							
R3	0.2	-0.17	1																						
R4	-0.6	-0.5	-0.1	1																					
HC	-0.1	0.3	-0.4	-0	1																				
HI	-0.1	-0.21	-0	0.09	-0.1	1																			
I	0.3	0.25	-0.2	-0.2	0.32	-0.46	1																		
LC	0.3	0.37	-0.4	-0.1	0.36	-0.09	0.4	1																	
LI	0.1	-0.12	-0.4	0.3	0.18	0.47	-0.1	0.24	1																
POP	0.7	0.56	0.04	-0.4	0.28	-0.14	0.4	0.46	0	1															
POPFAM	0.8	0.46	0.24	-0.4	0.06	-0.04	0.2	0.31	0.01	0.9	1														
AGEMED	0.7	0.5	-0.1	-0.5	0.48	-0.06	0.5	0.56	0.29	0.7	0.59	1													
AGEMEDM	0.7	0.51	-0.1	-0.5	0.48	-0.09	0.6	0.56	0.26	0.7	0.58	1	1												
AGEMEDF	0.7	0.5	-0.1	-0.5	0.49	-0.04	0.5	0.56	0.3	0.7	0.6	1	1	1											
VPH	0.7	0.53	0.27	-0.3	0.07	-0.3	0.3	0.32	-0.2	0.9	0.85	0.46	0.47	0.46	1										
AHHI	0.5	0.41	0.11	-0.4	0.29	-0.21	0.6	0.45	0.07	0.5	0.38	0.82	0.84	0.81	0.5	1									
LBF	0.6	0.59	0.01	-0.3	0.27	-0.22	0.4	0.46	-0.1	1	0.88	0.63	0.63	0.63	0.9	0.484	1								
UNEMR	0.3	0.27	-0.3	-0.3	0.53	0.14	0.5	0.49	0.44	0.5	0.34	0.85	0.84	0.86	0.1	0.557	0.4	1							
Lanes	0.3	0.2	0.04	-0.1	0	-0.17	0.1	0.11	-0	0.4	0.42	0.29	0.28	0.29	0.31	0.114	0.4	0.19	1						
Driveways	0.4	0.03	0.32	-0.3	-0.2	-0.14	0.4	0.14	-0	0.2	0.24	0.38	0.39	0.37	0.27	0.623	0.19	0.15	-0.1	1					
Signals	0	0.11	-0.1	0.33	0.31	0.2	0.1	0.53	0.34	0.4	0.3	0.24	0.23	0.25	0.3	0.156	0.4	0.25	-0	-0.05	1				
Median	0.2	-0.14	0.26	0.19	-0.5	0.11	-0	-0	0.01	0.1	0.25	-0.13	-0.1	0.17	-0.13	0.1	-0.2	0.28	0.13	0.112		1			
Horiz Curv	-0.1	-0.19	0.24	0.53	-0.2	0.04	-0.1	0.04	0.13	0.1	0.16	-0.2	-0.18	-0.2	0.26	-0.13	0.16	-0.2	-0	-0.02	0.334	0.2439	1		
Zone Length	0.3	0	0.26	0.38	-0.1	0.18	0	0.33	0.22	0.4	0.52	0.07	0.07	0.07	0.51	0.013	0.44	-0.1	0.14	0.06	0.592	0.4555	0.708	1	
AADT	-0.1	-0.14	0.08	0.41	0.19	0.15	-0.1	0.14	0.23	0.2	0.16	-0.1	-0.06	-0	0.15	-0.16	0.19	-0	0.26	-0.16	0.486	0.0925	0.531	0.608	1

Note: R1= 0.25 - 0.5 Acre Residential, R2= 0.25 Acre Residential/Apartments, R3=0.5 - 2 Acres Residential, R4= 2 Acres Residential/Open Space, HC=Heavy Commercial, HI=Heavy Industrial, I=Institutional,LC=Light Commercial, LI=Light Industrial, POP=Population , POPFAM=Population in Family Households, AGEMED=Median Age, AGEMEDM=Median Age Males, AGEMEDF=Median Age Females,VPH represents ownership, AHHI=Average household Income, UNEMR=Unemployment Rate, LBF Population In Labor Force, Lanes = Number of lanes, Driveways = Number of driveways, Signals= Number of signalized intersections, HorizCurv= Number of horizontal curvatures , Median= Median

ABLE8: Final Variables - 1.0-mile Buffer

Variables	Description
R1	0.25 - 0.5 Acre Residential
R3	0.5 - 2 Acre Residential
HI	Heavy Industrial
Median	Median
AADT	Average Annual Daily Traffic
Lanes	Number of lanes

A comparison of variables in tables 4, 6 and 8 indicates that land use variables (in particular, residential characteristics) seem to play a vital role when 0.25-mile buffer width is used. As buffer width increases, network characteristics seem to play a relatively higher role while the possible effect of land use characteristics seem to diminish. However, these differences observed seem to be more marginal in terms of selected variables to explain truck crash counts.

4.7 Generalized Linear Models

Crash estimation models were developed using the final set of variables identified from the correlation matrices using SPSS. GLM was used to analyze and develop these models. Both Poisson and NB log links were tested to select the distribution that better explains the relationship between dependent and independent variables. A Poisson model requires variance to mean ratio of the crash data to be about 1. When the assumption is not satisfied, over-dispersion occurs and the result will be a biased model.

In this research, SPSS was used to observe the variance-to-mean ratio in order to avoid over-dispersed or under-dispersed models. Poisson models were tested first. It was observed that the value/df of deviance and Pearson Chi-square statistics was observed to be greater than 1.0. This shows that the data is over-dispersed (i.e., variance is much greater than the mean). In order to minimize over-dispersion, NB distributions models were tested. Results obtained from these analyses are discussed in the following section.

4.7.1 Model based on 0.25-mile Buffer Width Dataset

Table9summarizes model outputs from analysis of 0.25-mile buffer width dataset. The P-value for all the variables shown in the table is less than 0.05. It can also be observed that coefficients (B) for the variables $0.25 - 0.5$ acre residential (R1), High Commercial (HC) and Industrial (I) are within the lower and upper limits of Wald confidence interval.

Overall, it was observed that High Commercial area (HC) and Industrial (I) have negative effect and 0.25 - 0.5 acre residential (R1) has a significant positive effect on the truck crash counts in buffers generated at 0.25-mile radius. The negative coefficient for Industrial (I) and High commercial (HC) indicates that the increase in industrial spaces and high commercial spaces rate results in a decrease of truck crash counts. Positive coefficient of 0.25-0.5 acre residential (R1) indicates that an increase in R1 type residential spaces shall increase truck crashes.

The final model developed using 0.25-mile buffer width dataset can be mathematically represented as

Log(crashes) = 3.639 + 0.002 * 0.25-0.5 acre residential (R1)− 0.005 *High Commercial(HC)− 0.025 * Industrial (I)

TABLE 9: Model Parameters Summary - 0.25-mile Buffer

Parameter	B	95% Wald Confidence Interval		
		Lower	Upper	P-Value
Intercept	3.639	3.171	4.106	-
R1	0.002	0.001	0.004	0.003
HC	-0.005	-0.008	-0.003	< 0.001
I	-0.025	-0.039	-0.012	< 0.001

4.7.2 Model based on 0.5-mile Buffer Width Dataset

Table10summarizes model outputs from analysis of 0.5-mile buffer width dataset. The P-values for all the variables in the final model (shown in the table) are less than 0.05. From the table, it can be observed that coefficients (B) for the variables 0.25 acre residential/apartments (R2), 0.5- 2 acre residential(R3) and Light Industrial (LI) are within the lower and upper limits of Wald confidence interval. In the model hence developed for 0.5-mile buffer width, it was observed that 0.25 acre residential/apartments (R2) and Light Industrial (LI)have significant effect on truck crash counts. The positive coefficient for 0.5-2 acres residential and Light Industrial(LI) area indicates that an increase in this variable results in increased truck crash counts. The negative coefficient for 0.25 acre residential/apartments (R2) indicates that an increase of R2 residential spaces decreases truck crash counts. An increase in light industrial space resulting in an increase in truck crash counts makes logical sense as industrial spaces attract more number of truck trips and can have a great effect on truck trips, hence justifying the

positive coefficient for light industrial spaces. The final developed based on 0.5-mile buffer width dataset can be mathematically represented

Log(Crashes) = 3.232 - 0.003 * 0.5 acres residential/apartments (R2) + 0.002 * 0.5 acres residential (R3)+0.002 * Light Industrial

TABLE10: Model Parameters Summary - 0.5-mile Buffer

Parameter	B	95% Wald Confidence Interval		
		Lower	Upper	P-Value
Intercept	3.232	2.8	3.663	-
R2	-0.003	-0.005	-0.001	0.001
R3	0.002	7.91E-05	0.005	0.043
LI	0.002	0	0.004	0.015

4.7.3 Model based on 1-mile Buffer Width Dataset

Table11summarizes model outputs from analysis of 1-mile buffer width dataset. It can also be seen that High Industrial area (HI) and presence of a median have a significant effect on truck crash counts at 95% confidence interval, as the corresponding P-values for these variables are less than 0.05. The coefficients (B) for these variables lie within the Wald confidence interval ranges. The positive coefficient for High Industrial (HI)area indicates that an increase in this variable will result in an increase in truck crash counts. The positive coefficient formedian shows that presence of medianincreases the truck crashes which is counter intuitive as median normally decreases crash counts and increases safety.

The final model developed using 1-mile buffer width dataset can be mathematically represented as

Log(crashes) = 2.827 + 0.001 * HI +0.424 * Median

TABLE11: Model Parameters Summary -1.0-mile Buffer

Parameter	B	95% Wald Confidence Interval		P-Value
		Lower	Upper	
Intercept	2.827	2.474	3.181	-
HI	0.001	0	0.002	0.013
Median	0.424	0.037	0.81	0.032

4.7.4 Summary and Selection of Best Model / Spatial Proximity

As discussed earlier in Chapter 3, Deviance, degrees of freedom (df),Log Likelihood, Pearson Chi-Square, AIC and AICC statistics are computed to select the best model among the models generated. In general, the crash estimation model with lower AICC value and the value of deviance/df close to one is selected as the best model. The goodness of fit test values for each buffer width data set are summarized in Table12.

TABLE 12:Summary of Goodness of Fit Statistics

Goodness of Fit Statistics									
Criterions	0.25 mile Buffer			0.5 mile Buffer			1.0 mile buffer		
	Value	df	Value/df	Value	df	Value/df	Value	df	Value/df
Deviance	25.374	19	1.335	44.287	35	1.265	59	50	1.18
Pearson Chi-Square	17.086	19	0.899	31.182	35	0.891	44.236	50	0.885
Log Likelihood	-88.55			-165.96			-225.9		
Akaike's Information Criterion (AIC)	187.1			341.915			459.77		
Finite Sample Corrected AIC (AICC)	190.44			343.68			460.58		

An increase in computed AIC and AICC was observed with an increase in buffer width to extract off-network characteristics data and develop truck crash estimation models.

Table 13 summarizes the observed effect of independent variables on truck crash counts in the models for each buffer width dataset. From the table, it can be observed that 0.25-0.5 acres residential (R1), 0.5 acres residential (R3), Light Industrial (LI), High Industrial (HI), Median have a positive influence on truck crashes implying that an increase in the number/quantity of these variables will result in an increase in the number of truck crashes. Variables such as0.5 acres residential/apartments (R2), and High Industrial (HI)have a negative effect on truck crash counts (as these variables increase the number of truck crashes decrease and vice versa).This makes sense as residential areas might generate or attract relatively less number of truck trips. High commercial and Industrial areas have a negative effect on the truck crashes. In general, commercial and industrial areas influence truck crashes. This could be because a warehouse or an industry would attract more number of truck trips than a residential area or educational institute. But the increase of industrial space decreasing the number of truck crashes in 0.25 mile buffer zone is counter-intuitive. Similarly, median having a positive effect in case of 1.0-mile buffer width model does not make much sense as generally median reduces head on collisions and other type of crashes aiding in enhancing safety. However, issues such as limited turning radius might increase truck crash counts in the presence of median.

TABLE13: Influence of Independent Variables

Variables	0.25 Mile Buffer	0.5 Mile Buffer	1.0 Mile Buffer
R1	+		
R2		-	
R3		+	
HC	-		
LI		+	
HI			+
I	-		
Median			+

The value/df is found to be reasonably close to one for all three buffer widths. The AICC was observed to be the lowest for 0.25–mile buffer width dataset. But the overall model for 0.25 mile buffer width was not statistically significant. Also, the signs for coefficients of variables are not meaningful (counter-intuitive). Hence the NB model based on 0.5-mile buffer width dataset could be considered as the statistically best model to estimate truck crashes on intra city roads.

4.8 Validation

Models were validated by substituting data for eight randomly selected crash zones that were not used for model development. Models selected were validated by substituting relevant values in the mathematical regression equations that were formulated for each buffer width and compared to the actual truck crash counts. The percentage of error between the two was then calculated. The difference between the

crashes computed and crashes in the crash data available was observed to range between %-90 to 56 % for 0.25-mile regression model.

Models developed for 0.5-mile buffer width and 1.0-mile buffer width are also validated to see the percent difference between actual truck crash counts and predicted truck crash counts. The percent difference varied from –5% and 7% in case of 0.5-mile buffer width scenario whereas between -19% and 16% for 1.0-mile buffer width scenario. Quite some difference can be observed between actual truck crash counts and statistically predicted counts for the latter two buffer widths as well. Table 14 summarizes the actual truck crash counts and computed truck crash count values for the eight zones that were not used for model development.

In addition, the root mean square error (RMSE), which measures the average of the squares of the errors, was also calculated. The computed RMSE was found to be the lowest (9.72 %) for 0.5-mile buffer width dataset model and the highest (55.09 %) for 0.25-mile buffer width dataset model. Thus, results from validation also indicate that 0.5-mile buffer width dataset model gives results relatively closer to real-world observations.

TABLE 14: Validation of Models

Zone Number	Log(Actual Counts)	0.25 Mile		0.5 Mile		1.0 Mile	
		Log (Calculated Crash Counts)	Percent Difference	Log (Calculated Crash Counts)	Percent Difference	Log (Calculated Crash Counts)	Percent Difference
1	0.60	0.75	-24.57	0.61	-0.54	0.51	16.07
2	1.15	1.48	-29.13	1.18	-2.53	1.37	-19.53
3	0.85	0.46	46.04	0.85	-0.92	0.76	10.36
4	1.43	1.77	-23.31	1.42	1.07	1.52	-5.91
5	1.28	1.82	-42.37	1.19	7.04	1.49	-16.32
6	1.04	1.99	-90.75	1.08	-4.09	1.18	-13.69
7	0.48	0.87	-83.18	0.45	4.94	0.57	-19.44
8	0.60	0.26	56.65	0.61	-5.28	0.70	-17.01
Root Mean Square Error		55.09		9.72		15.37	

CHAPTER 5: CONCLUSIONS

This research presents a geospatial method to identify high dense truck crash zones and develop intracity truck crash estimation models. Data for the city of Charlotte, North Carolina for the year 2008was used to illustrate the geospatial method and develop the model. Analyses showed that most of the truck crashes on intracity roads were found to be injury type or property damage only crashes.

High truck crash zones were identified using Kernel density method and were ranked using several parameters. Kernel density method helped in identifying the hot spot pockets of truck crashes and sixty linear zones were picked from the density map to study the effect of various on- and off-network characteristics around each zone on truck crashes.

Demographic, socio-economic, network and land use characteristics were used as predictor variables. Variables from these several datasets were used in developing correlation matrices and generalized linear models.

Data were extracted using 0.25-, 0.5-and 1-mile buffer widths to study the role of spatial proximity in capturing demographic, socio-economic and land use data for developing models. Based on the goodness of fit tests, it was observed that the model obtained by using 0.5-mile buffer width is the best model when compared to other buffer widths.

The results obtained are also supported by findings from validation of developed models. The methods proposed in this research are easy to adopt and can be applied universally to urban settings of any size and level. The method developed for identification of high truck crashes zones can be implemented in any region to know the hot spot areas with regard to truck crashes. It can help identify where truck crashes are high and allocate resources based on rank (higher the risk, lower the rank).The method and models developed from this research can be proactively used to estimate intracity truck crashes, incorporate them into planning process and enhance safety. While not many changes can be made towards socio-economic and demographic characteristics, certainly strict land use regulations and altering on-network characteristics can be promoted by transportation agencies to ensure more safety with regard to truck crashes.

Overall the findings from this research with abilities to infer meaning from complicated data seems useful in

- providing robust understanding of truck crash patterns on intracity roads;

- aiding Metropolitan Planning Organizations (MPOs) to proactively incorporate truck crashes and their effects in planning process;

- incorporate safety planning at micro levels; and

- assist in categorizing truck routes for congestion management planning.

This research at broader level enhances truck safety and other system users by formulating a systematic methodology. While the identification of the high-risk zones is critical for effective allocation of resources, the methodology developed to estimate intra city truck crashes is vital to conduct micro level traffic safety and selection of suitable countermeasures. Though data such as population by age group and temporal variations

of truck crashes were not available and so could not be used in this research. Developing models based on variables that are more critical can help develop better crash estimation models.

The effect of land use characteristics on crashes emphasizes the need to further explore their role on mobility and safety. It also stresses the need to develop truck trip generation and planning models.

REFERENCES

Abdel Aty, M. A., P.Anurag, L. Chris, G.Vikas, S. Dos. 2007. "Crash risk assessment using intelligent transportation systems data and real-time intervention strategies to improve safety on freeways".Journal of Intelligent Transportation Systems, 11(3), pp. 107-120.

Agarwal, R., and D. Lord. 2006. "Effect of sample size on goodness-of-fit statistics and confidence interval models subjected to low sample mean values". Transportation Research Board, pp. 35-43.

Abdel Wahab, H.T., and M.A.Abdel-Aty. 2001. "Development of artificial neural network models to predict driver injury severity in traffic accidents at signalized intersections". Transportation Research Record 1746, pp. 6-13.

Adams,B. 1989."Deregulation's negative effect on safety". In: Moses, L.N., Savage, I. (Eds.) Transportation Safety in an Age of Deregulation. Oxford University Press, New York, pp. 21–27.

Alexander, D.L. 1992. "Motor carrier deregulation and highway safety: an empirical analysis". Southern Economic Journal 59, (1), pp. 28–38.

Alassar, L. 1988. "Analysis of heavy truck accident severity". Journal of Advanced Transportation, 22, pp. 77-91.

Bedard, M., G. H. Guyatt, M. J. Stones, and J.P. Hirdes. 2002. "The independent contribution of driver, crash, and vehicle characteristics to driver fatalities". Accident Analysis and Prevention, 34 (6), pp. 717–727.

Berkowitz,A. 2001. "The marriage of safety and land-use planning: A fresh look at local roadways".FHWA, Journal of Public Roads, 65.

Bindra, S, I. N. John, and J. Thomas. 2009. "Predicting segment-intersection crashes with land development data".Transportation Research Record 2102,pp.9-17.

Campbell, K.L. 1991. "Fatal accident involvement rates by driver age for large trucks". Accident Analysis and Prevention, 23 (4), pp. 287–295.

Chang, L.Y., and F. Mannering. 1999. "Analysis of injury severity and vehicle occupancy in truck- and non-truck-involved accidents". Accident Analysis and Prevention, 31, pp. 579–592.

Chira-Chavala, T. et al. 1984. "Severity of large-truck and combination vehicle accidents in over-the road service: A discrete multivariate analysis". Transportation Research Record 975, pp. 23-56.

Craft, R. 1999. "Longer combination vehicles involved in fatal crashes 1991–1996".Federal Highway Administration, Office of Motor Carrier Research and Standards, Washington DC.

Daniel., J., and S. Chien. 2003. "Identifying factors and mitigation technologies in truck crashes in New Jersey".National Centre for Transportation and Industrial Productivity, New Jersey Institute of Technology.

Debauche, W., and Decock, D. 2007. "Working group on longer and heavier goods vehicles (LHVs): Multidisciplinary approach to the issue)".Belgian Road Research Centre Appendix to the BRRC Bulletin Number 70, Brussels, Belgium.

Dobbs, G. L. 2009. "Pedestrian and bicycle safety on acollege campus". Crash and Conflict Analysis With Recommended Design.Alternatives for Clemson University, Master's Thesis.

Duncan, C.S., A. J. Khattak, and F. M. Council. 1998. "Applying the ordered-probit model to injury severity in truck-passenger car rear-ends collisions". Transportation Research Record 1635, pp. 63–71.

Eluru, N., and C. Bhat.2007. "A joint econometric analysis of seat belt use and crash-related injury severity". Accident Analysis and Prevention, 39(5), pp. 1037-1049.

Evans, L.,and M.C. Frick. 1993. "Mass ratio and relative driver fatality risk in two-vehicle crashes". Accident Analysis and Prevention, 25 (2), pp. 213–224.

Garber, N. J. 2001. "Stochastic models relating crash probabilities with geometric and corresponding traffic characteristics data".Research Report No.UVACTS-5-15-74.

Glaeser, K.P., and R. Kaschner. 2006. "Effects of new vehicle concepts on the infrastructure of the federal trunk road network".Final report, Federal Highway Research Institute, Berlin, Germany.

Golob, T.F., and W.W. Recker. 2003. "Relationships among urban freeway accidents, traffic flow, weather, and lighting conditions". Journal of Transportation Engineering, 129(4), pp. 342-353.

Golob, T. et al. 1987."An analysis of the severity and incident duration of truck-involved freeway accidents".Accident Analysis and Prevention, 19(5), pp 375-375.

Gross, F., P. P. Jovanis, K. Eccles, and K.Y. Chen. 2009. "Safety evaluation of lane and shoulder width combinations on rural, two-lane, undivided Roads". Report No. FHWA-HRT-09-031.U.S.Department of Transportation, Federal Highway Administration, Turner-Fairbank Highway Research Centre.

Hanley P.F., and D. J. Forkenbrock. 2005. "Safety of passing longer combination vehicles on Two-lane Highways". Transportation Research Part A, 39, pp. 1–15.

Hauer, E., J. Kononov, B. Allery, and M. S. Griffith. 2002. "Screening the road network for sites with promise". Transportation Research Record 1784, pp.27-32.

Hauer, E., J. C. N. Ng, and J. Lovell. 1998. "Estimation of safety at signalized intersections". Transportation Research Record 1185, pp. 48-61.

Heimbach, C.L., W. W. Hunter, and G.C. Chao. 1974. "Paved highway shoulders and accidentexperience". Journal of Transportation Engineering, Vol. 100(4), pp. 887-908.

Huelke, D. F., and C.P. Compton. 1995. "The effects of seat belts on injury severity of front and rear seat occupants in the same frontal crash". Accident Analysis and Prevention, 27 (6), pp. 835-838.

Ivan, J.N., R. K. Pasupathy, and P. J. Ossenbruggen. 1999. "Differences in causality factors for single and multi-vehicle crashes on two-lane roads". Accident Analysis and Prevention, 31(6), pp. 695-704.

Jones, A.P., and S.H. Jorgensen. 2003. "The use of multilevel models for the prediction of road accident outcomes". Accident Analysis and Prevention, 35(1), pp. 59-61.

Jovanis, P. P.m and H. L. Chang. 1986. "Modelling the relationship of accidents to miles travelled". Transportation Research Record 1068,pp42-51.

Khattak, A.J., R. J. Schneider, and F.Targa. 2003. "Risk factors in large truck rollovers and injury severity: analysis of single-vehicle collisions". Presented at 2003 Annual Transportation Research Board Meeting, CD-ROM, Washington, DC.

Khasnabis, S. et al. 1989. "Use of time series analysis to forecast truck accidents".In Transportation Research Record 1249, pp 30-36.

Knight, I., W. Newton, and McKinnon, et al. YEAR "Longer and/or Longer and Heavier Goods Vehicles (LHVs): A Study of the Likely Effects if Permitted in the UK".Final Report PPR Number 285, TRL Limited, Berkshire, United Kingdom.

Knipling, R.R. 2008 "Comparison of combination-unit and single-unit trucks in the large truck crash causation study". Federal Motor Carrier Safety Administration Webinar, Virginia Tech Transportation Institute.

Kraas, A.1993. "The impact of the US motor carrier act 1980 on road safety".In California: an econometric policy evaluation Logistics and Transportation Review 29, pp. 179–192.

Kiattikomol V., C. Arun, J. E. Hummer, and S. M. Younger. 2008. "Planning level regression models for prediction of crashes on interchange and no interchangesegments of urban freeways". Journal of Transportation Engineering, 134(3), pp. 111-117.

Kim, D., Y. Lee, S. P. Washington, and K. Choi. 2006. "Modeling crash outcome probabilities at rural intersections: Application of hierarchical binomial logistic models". Accident Analysis and Prevention, 39(1), pp. 125-134

Kim, K., B. Made, and Y. Eric. 2006. "Influence of land Use, population, employment, and economic activity on accidents".Journal of Transportation Research Board1953, pp. 56-64.

Lee, J., and F. Mannering. 2002. "Impact of roadside features on the frequency and severity of run-off-roadway accidents: An empirical analysis". Accident Analysis and Prevention, 34(2), pp. 149-161.

Lemp, D. J., K. M. Kara and U. Avinash. 2010. "Analysis of large truck crash severity using heteroskedastic ordered-probit models". Accident Analysis and Prevention, 43,pp.370–380.

Liesman, J. S. 2005 "An analysis of Wyoming truck crashes".Department of Civil and Architecture Engineering,University of Wyoming,Master's Thesis.

Lord, D., S. P. Washington, and J. N. Ivan. 2005. "Poisson, Poisson-gamma and zero inflated regression models of motor vehicle crashes: balancing statistical fit and theory". Accident Analysis and Prevention, 37 (1), pp35-46.

Gord., L. G. L. Clark., and Sayed T. 2003 "Community-based, macro level collision prediction model use with a regional transportation plan". Journal Of Transportation Engineering. Canadian Journal of Civil Engineering, 30(4) pp. 711-719.

Maycock,G., and R. D. Hall. 1984. "Accidents at 4-am roundabouts". Report 1120 Crowthorne, U.K.: Transport and Road Research Laboratory 1984.

Singh, M. 2009. "A statistical model of fatality rates in large truck crashes".North Dakota State University, Department of Agriculture and Applied Science, Master's Thesis.

Miaou, S. P., and D. Lord. (2003). "Modeling traffic crash flow relationships for intersections: Dispersion parameter, functional form, and bayes versus Empirical Bayes methods". Transportation Research Record 1840, pp. 31-40.

Miaou, S. P. (1994). "The relationship between truck accidents and geometric design of road section: Poisson versus Negative Binomial regression". Accident Analysis & Prevention, 26(4), pp. 471-482.

Miller, J. S. 1999. "What value may Geographic Information Systems add to the art of identifying crash countermeasures?".Report 99-R13,Virginia Transportation Research Council.April 1999.

Mitra.,S. and S. Washington. 2006. "On the nature of over-dispersion in motor vehicle crash prediction models".Civil & Environmental Engineering Department, Cal Poly State University, San Luis Obispo, USA.

Moore, T.G. 1989. "The myth of deregulation's negative effect on safety". In: Moses, L.N., Savage, I. (Eds.) Transportation Safety in an Age of Deregulation. Oxford University Press, New York, pp. 8–20.

NassiriH.,and A. Edrissi.2006, "Modeling truck accident severity on two-lane rural highways".ScientiaIranica, Sharif University of Technology,Vol. 13, No. 2, pp 193 -200

NHTSA. 2006. "Traffic safety facts 2005: Large trucks". Final Report # DOT HS 810 619, http://wwwnrd.nhtsa.dot.gov/Pubs/810619.PDF.

NHTSA. 2008. "Traffic safety facts 2007: Large trucks". Final Report # DOT HS 810 805.http://wwwnrd.nhtsa.dot.gov/Pubs/810989.PDF.

Nukoolkit, C. and H. C. Chen.2001. "A data transformation technique for car injury prediction".Computer Science Department, University of Alabama, Master'sThesis.

Okamoto, H. andM. Koshi.1989. "A method to cope with the random errors of observed accident rates in regression analysis". Accident Analysis and Prevention Journal, 21, pp. 317-332.

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 and Prevention, 28 (6), pp. 739-753.

Park, B. J. 2010. "Application of finite mixture models for vehicle crash data analysis".Department of Civil Engineering, Texas A&M University, College Station, TX 77843-3136.

Peltzman, S. 1975. "The effects of automobile safety regulations". Journal of Political Economy, 83 (4), 677–725.

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

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

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

Pulugurtha, S. S., K. K.Vanjeeswaran, and M.Uddaraju. 2004. "Development of criteria to identify pedestrian high crash locations in Nevada". Report # RDT 04-048; Nevada Department of Transportation. Carson City, NV.

Renshaw, N.2007."Longer and heavier lorries (LHLs) and the environment".Position Paper European Federation for Transport and Environment,96/53/EC, 2008.

Roy Jorgensen Associates Inc. 1978."Cost and safety effectiveness of highway design elements".Report 197, Transportation Research Board, National Research Council, Washington, D.C., 1978.

Thobias, S. 2005. "Modelling highway crashes using Bayesian belief networks technique and GIS".College of Engineering,The Florida State University,PhD Dissertation.

Savolainen, P., and F. Mannering. 2007. "Probabilistic models of motorcyclists injury severities in single- and multi-vehicle crashes". Accident Analysis and Prevention, 39(5), pp. 955-963.

Song J.J., M. Ghosh, S. .Miaou, and B.Mallick. 2004. "Bayesian multivariate spatial models for roadway traffic crash mapping".Department of Statistics, TAMU 3143, Texas A&M University, College Station, TX 77843-3143, USA.

Srinivasan.,K.K.. 2002. "Injury severity analysis with variables and correlated thresholds: Ordered mixed logitformulation".Transportation Research Record 1784. pp. 132-142.

Saccomanno, F.F. 1988. "Generalized log linear models of truck accident rates".In Transportation Research Record 1172 pp. 23-31.

Son.,D. H., et al., 2011. "Development of crash prediction models with individual vehicular data".Transportation Research Part C: Emerging technologies, 19, pp. 1353-1363.

Ulfarsson, G.F., and F. Mannering. 2004. "Differences in male and female injury severities in sport-utility vehicle, minivan, pickup and passenger car accidents". Accident Analysis and Prevention, 36(2), pp. 135-147.

US Department of Transportation – USDOT. 2000. "Comprehensive truck size and weight study, Volume III, scenario analysis".US Department of Transportation, Washington, DC.

US Government Accounting Office (GAO). 1992. "Truck safety: The safety of longer combination vehicles is unknown". Report RECD-92-66 Government Accounting Office, Washington, DC.

Vavilikolanu, S. 2008. "Crash prediction models on truck-related crashes on two-lane rural highways with vertical curves".The University of Akron, Master'sThesis.

Wang, X.K., and K.M. Kockelman. 2005. "Use of heteroscedasticordered logitmodel to study severity of occupant injury". Transportation Research Record 1908, pp. 195-204.

Wang, J.S., R. R. Knipling, and L. J., Blincoe. 1999. "The dimensions of motor vehicle crash risk". Journal of Transportation and Statistics, 2 (1), 19–43.

Washington, S., J. Metarko, and I.Fomunung.et al., 1999 "An inter-regional comparison: Fatal crashes in the south-eastern and non-south-eastern United States: Preliminary findings". Accident Analysis and Prevention, 31(1), pp. 135-146.

Wang, Y., and N. L. Nihan.2003). "Quantitative analysis on angle-accident risk at signalized intersections".Accident Analysis and Prevention,36,pp. 313-321.

Wood, G. R. (2002). "Generalized linear accident models and goodness of fit Testing". Accident Analysis & Prevention, 34(4), pp. 417-427.

Wood, G. R. (2005). "Confidence and prediction intervals for generalized linear accident models". Accident Analysis & Prevention, 37(2), pp. 267-273.

Xie.,Y, L.Dominique, and Y. Zhan. 2006. "Predicting motor vehicle collisions using Bayesian neural network models: An empirical analysis".Final Report, Zachry Department of Civil Engineering, Texas A&M University.

Yau, K.K.W., H. P. Lo, and S.H.H. Fung. 2006. "Multiple-vehicle traffic accidents in Hong Kong".Accident Analysis and Prevention, 38(6), pp.1157–1161.

Zegeer, C. V.,R. Stewart, D.,Reinfurt, F. Council, T.Neuman, E Hamilton, T. Miller, and W. Hunter, W. 1990. "Cost effective geometric improvements for safety upgrading of horizontal curves".Federal Highway Administration, Publication No.FHWA-RD-87-008, Washington, DC.

Zhu., X and S. Sivaramakrishnan. 2011. "A comprehensive analysis of factors influencing the injury severity of large-truck crashes" Accident Analysis and Prevention, 43, pp. 49-57.